Generated on 28/11/2026 21:43

KNOWLEDGE BASE · PUBLIC DOCUMENT

Context and output limits in OpenCode: limit.context and limit.output

OpenCode

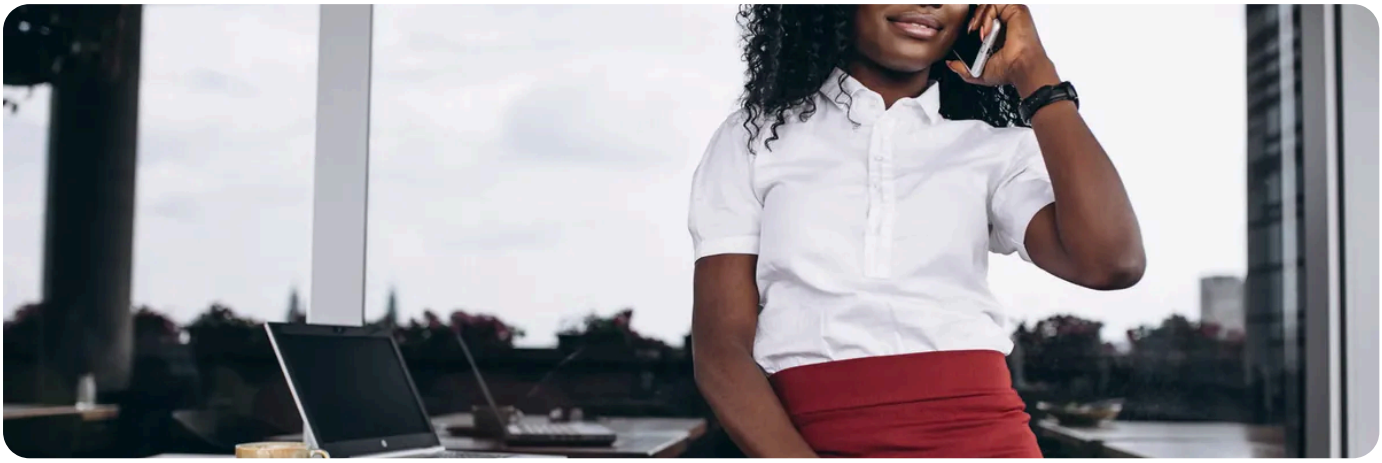
CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-9660	v2026.11.28	International · English

MozOut International
RLX Holdings (Pty) Ltd. · Group RLX Holdings
Tax no. (SARS) 943 907 6234 · Registration no. (CIPC) 2025/667953/07

https://www.my.mozout.com/knowledgebase/9660

Context and output limits in OpenCode: limit.context and limit.output

OpenCode Created on **November 28, 2026** 2 min read



Each model in the `models` block can have a `limit` with two numbers: the context and the output. OpenCode does not know our models, so it is worth giving it these values. They change when it summarises the conversation and how long each answer can be.

The two numbers

```
"moonshotai/kimi-k2.7-code": { "name": "Kimi K2.7 Code", "limit": { "context": 262144, "output": 16000 } }
```

Field	What to put	Where to see it
`limit.context`	The model's context: what it reads in one request, input and output together.	`context_length` in `GET /v1/models`
`limit.output`	The most tokens per answer. Our examples use `16000`.	Do not go over the model's `max_output_tokens`.

What OpenCode does with them

When the conversation gets close to the context limit, OpenCode summarises the older part and carries on. To do that, it needs to know the model's real context.

Without `limit`, according to its documentation, OpenCode assumes 200,000 tokens of context and 32,000 of output for a model it does not know. The values you write replace these.

If the context in the file is larger than the real one, the conversation can go over the model's limit before OpenCode summarises it. If it is smaller, it summarises earlier than it needs to.

Why 16,000 of 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. Before each answer, we reserve credits for the input and for the maximum output. If the balance does not cover the reservation, the API answers `402 insufficient_credits`. At the end you only pay for the real tokens, and the rest of the reservation goes back to your balance.

If you need longer answers, raise `output`, without going over the model's `max_output_tokens`.

The values for the recommended models

Model	limit.context	limit.output
`moonshotai/kimi-k2.7-code`	`262144`	`16000`
`deepseek-ai/deepseek-v4-pro`	`1048576`	`16000`
`zai-org/glm-5.2`	`1048576`	`16000`
`deepseek-ai/deepseek-v4-flash`	`1048576`	`16000`

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

```
curl "https://api.mozout.com/v1/models?type=text" -H "Authorization: Bearer $MOZOUT_API_KEY"
```

In Windows PowerShell:

```
curl.exe "https://api.mozout.com/v1/models?type=text" -H "Authorization: Bearer $env:MOZOUT_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.

The block in **OpenCode with All in One AI** already has the `limit` of each of the 79 models that support tools.

RECOMMENDED

Want OpenCode to spend only what it needs? Visit the Coding with AI page.

See Coding with AI →

SEE ALSO

- OpenCode with All in One AI: creating opencode.json from scratch, the key and the models
- Adding more models to OpenCode: the models block and the id with the slash
- OpenCode and the Coding Plan: points, credits and what is left
- Context and output limits in kilo.jsonc
- Coding with AI in All in One AI

Mozout

MozOut International

RLX Holdings (Pty) Ltd. · Group RLX Holdings

TAX NO. (SARS)	943 907 6234
REGISTRATION NO. (CIPC)	2025/667953/07
REGISTERED OFFICE	1111 Olive Street Saint Louis 63101, Missouri, United States
CONTACT	info@mozout.com · +27 78 823 3325 · www.mozout.com
OPENING HOURS	Monday to Friday - 08 AM to 04 PM Saturday & Sunday - Closed

Online version of this document

RLX-INT-KB-9660 · v2026.11.28 · Generated on 28/11/2026 21:43