Mozout

Generated on 28/11/2026 21:45

KNOWLEDGE BASE · PUBLIC DOCUMENT

Context and output limits in kilo.jsonc

Kilo Code

CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-9708	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/9708

Context and output limits in kilo.jsonc

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



Each model in `kilo.jsonc` can have a `limit` with the context and the output. For your own providers, the Kilo Code documentation says to always set both. Without them, Kilo Code does not summarise the conversation and asks for long answers.

The two fields

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

Field	What it is	What to put
`limit.context`	The model's total context. Kilo Code uses it to know when to summarise the conversation.	The model's `context_length`, in `GET /v1/models`.
`limit.output`	The most tokens per answer. Kilo Code sends it to the API as `max_tokens`.	`16000`, without going over the model's `max_output_tokens`.

There is also `limit.input`, optional: a lower ceiling for the input. If you set it, Kilo Code summarises the conversation against it instead of `context`.

Without limit

Kilo Code looks first in your file and then in its own model catalogue. For a model that is not in that catalogue, the values are 0:

Missing value	What happens
`context` at 0	Kilo Code does not summarise the conversation. It grows until the API rejects the request.
`output` at 0	Kilo Code uses its own default, 32,000 tokens.

Why the output matters for your credits

Before each answer, we reserve credits for the input and for the `max_tokens` that Kilo Code sends. With 32,000, the output part of the reservation is twice that of 16,000, and with a low balance 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.

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.

In the VS Code extension too

The **Custom provider** window has no fields for the limits. According to the Kilo Code documentation, the token limits, tools and variants of your own provider are set in `kilo.jsonc`. The full file is in the article on **the global kilo.jsonc**.

RECOMMENDED

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

See Coding with AI →

SEE ALSO

- Kilo Code with All in One AI: the custom provider in VS Code and kilo.jsonc in the CLI
- Kilo Code on the command line: the global kilo.jsonc file
- Choosing a model for Kilo Code: models that accept tools
- Context and output limits in OpenCode: limit.context and limit.output
- 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-9708 · v2026.11.28 · Generated on 28/11/2026 21:45