interweb

Gerado a 28/11/2026 20:48

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Limites de contexto e de resposta no kilo.jsonc

Kilo Code

CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-AO-KB-3208	v2026.11.28	Angola · Português

Interweb Angola
RLX Angola, Lda. · Grupo RLX Holdings
NIF 500 260 5724 · Matrícula 26464-25/250724

https://www.meu.interweb.ao/knowledgebase/3208

Limites de contexto e de resposta no kilo.jsonc

Kilo Code Criado em **28 de novembro de 2026** 2 min de leitura



Cada modelo do `kilo.jsonc` pode ter um `limit` com o contexto e a saída. Para fornecedores próprios, a documentação do Kilo Code manda pôr sempre os dois. Sem eles, o Kilo Code não resume a conversa e pede respostas longas.

Os dois campos

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

Campo	O que é	O que pôr
`limit.context`	O contexto total do modelo. O Kilo Code usa-o para saber quando resumir a conversa.	O `context_length` do modelo, em `GET /v1/models`.
`limit.output`	O máximo de tokens de cada resposta. O Kilo Code manda-o à API como `max_tokens`.	`16000`, sem passar o `max_output_tokens` do modelo.

Há ainda o `limit.input`, opcional: um tecto mais baixo para a entrada. Se o puser, o Kilo Code resume a conversa por ele, em vez do `context`.

Sem limit

O Kilo Code procura primeiro no seu ficheiro e depois no catálogo de modelos dele. Para um modelo que não está nesse catálogo, os valores ficam a 0:

Valor em falta	O que acontece
`context` a 0	O Kilo Code não resume a conversa. Ela cresce até a API recusar o pedido.
`output` a 0	O Kilo Code usa o valor dele por omissão, 32 000 tokens.

Porque a saída conta para os créditos

Antes de cada resposta, reservamos créditos para a entrada e para o `max_tokens` que o Kilo Code manda. Com 32 000, a reserva da saída é o dobro da de 16 000, e com pouco saldo a API responde `402 insufficient_credits`. No fim paga só os tokens verdadeiros, e o resto da reserva volta ao saldo.

Os valores dos modelos recomendados

Modelo	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`

Os ids estão na aba «Modelos» do **All in One AI** ou na lista da API. No Linux e no macOS:

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

No PowerShell do Windows:

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

Cada modelo vem com o `id` (sempre com a barra), o `context_length` e o preço em `pricing`. Ver a lista não gasta créditos.

Também na extensão do VS Code

A janela **Custom provider** não tem campos para os limites. Segundo a documentação do Kilo Code, os limites de tokens, as ferramentas e as variantes de um fornecedor próprio acertam-se no `kilo.jsonc`. O ficheiro completo está no artigo sobre **o kilo.jsonc global**.

RECOMENDADO

Quer o Kilo Code a gastar só o que precisa? Veja a página Programar com IA.

Ver Programar com IA →

VEJA TAMBÉM

- **Kilo Code com o All in One AI: o fornecedor próprio no VS Code e o kilo.jsonc na linha de comandos**
- **O Kilo Code na linha de comandos: o ficheiro kilo.jsonc global**
- **Escolher um modelo para o Kilo Code: modelos que aceitam ferramentas**
- **Limites de contexto e de resposta no OpenCode: limit.context e limit.output**
- **Programar com IA no All in One AI**

interweb

Interweb Angola

RLX Angola, Lda. · Grupo RLX Holdings

NIF	500 260 5724
MATRÍCULA	26464-25/250724
SEDE	Rua Direita do Patriota, Bairro Dangereux Belas, Luanda, Angola
CONTACTOS	info@interweb.ao · +244 942 011 191 · www.interweb.ao
HORÁRIO	2ª. à 6ª. Feira - 08 à 16 Horas Sábado & Domingo - Fechado

Versão online deste documento

RLX-AO-KB-3208 · v2026.11.28 · Gerado a 28/11/2026 20:48