interweb

Gerado a 28/11/2026 20:48

BASE DE CONHECIMENTO · DOCUMENTO PÚBLICO

Limites de contexto e de resposta no OpenCode: limit.context e limit.output

OpenCode

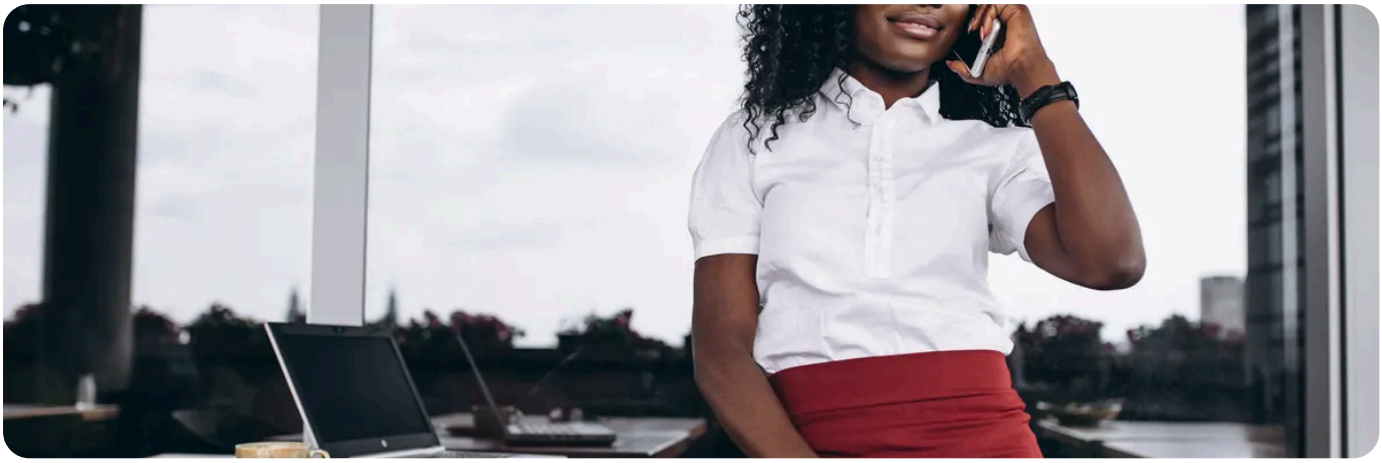
CÓDIGO	VERSÃO	PAÍS · IDIOMA
RLX-AO-KB-3192	v2026.11.28	Angola · Português

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

https://www.meu.interweb.ao/knowledgebase/3192

Limites de contexto e de resposta no OpenCode: limit.context e limit.output

OpenCode Criado em **28 de novembro de 2026** 2 min de leitura



Cada modelo do bloco `models` pode ter um `limit` com dois números: o contexto e a saída. O OpenCode não conhece os nossos modelos, por isso vale a pena dar-lhe estes valores. Mudam a altura em que ele resume a conversa e o tamanho que cada resposta pode ter.

Os dois números

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

Campo	O que pôr	Onde o ver
`limit.context`	O contexto do modelo: o que ele lê num pedido, entrada e saída juntas.	`context_length` em `GET /v1/models`
`limit.output`	O máximo de tokens de cada resposta. Os nossos exemplos usam `16000`.	Não passe o `max_output_tokens` do modelo.

O que o OpenCode faz com eles

Quando a conversa se aproxima do limite de contexto, o OpenCode resume a parte mais antiga e continua. Para isso precisa de saber o contexto verdadeiro do modelo.

Sem `limit`, segundo a documentação dele, o OpenCode assume para um modelo que não conhece 200 000 tokens de contexto e 32 000 de saída. Os valores que escrever substituem estes.

Se o contexto do ficheiro for maior do que o verdadeiro, a conversa pode passar o limite do modelo antes de o OpenCode a resumir. Se for menor, ele resume mais cedo do que precisava.

Porquê 16 000 de saída

Reservamos créditos pelo máximo de saída que a ferramenta pede em cada resposta. Por isso os exemplos pedem 16 000 tokens: chega para programar e precisa de menos saldo do que um valor maior. Antes de cada resposta, reservamos créditos para a entrada e para a saída máxima. Se o saldo não chega para a reserva, a API responde `402 insufficient_credits`. No fim paga só os tokens verdadeiros, e o resto da reserva volta ao saldo.

Se precisar de respostas maiores, suba o `output`, sem passar o `max_output_tokens` do modelo.

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.

O bloco de **OpenCode com o All in One AI** já traz o `limit` de cada um dos 79 modelos que aceitam ferramentas.

RECOMENDADO

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

Ver Programar com IA →

VEJA TAMBÉM

- OpenCode com o All in One AI: criar o opencode.json do zero, a chave e os modelos
- Juntar mais modelos ao OpenCode: o bloco models e o id com a barra
- O OpenCode e o Coding Plan: pontos, créditos e o que resta
- Limites de contexto e de resposta no kilo.jsonc
- 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-3192 · v2026.11.28 · Gerado a 28/11/2026 20:48