Mozout

Generated on 28/11/2026 21:53

KNOWLEDGE BASE · PUBLIC DOCUMENT

The recommended coding models: the context and what each one is for

Choosing a Tool & Fixing Errors

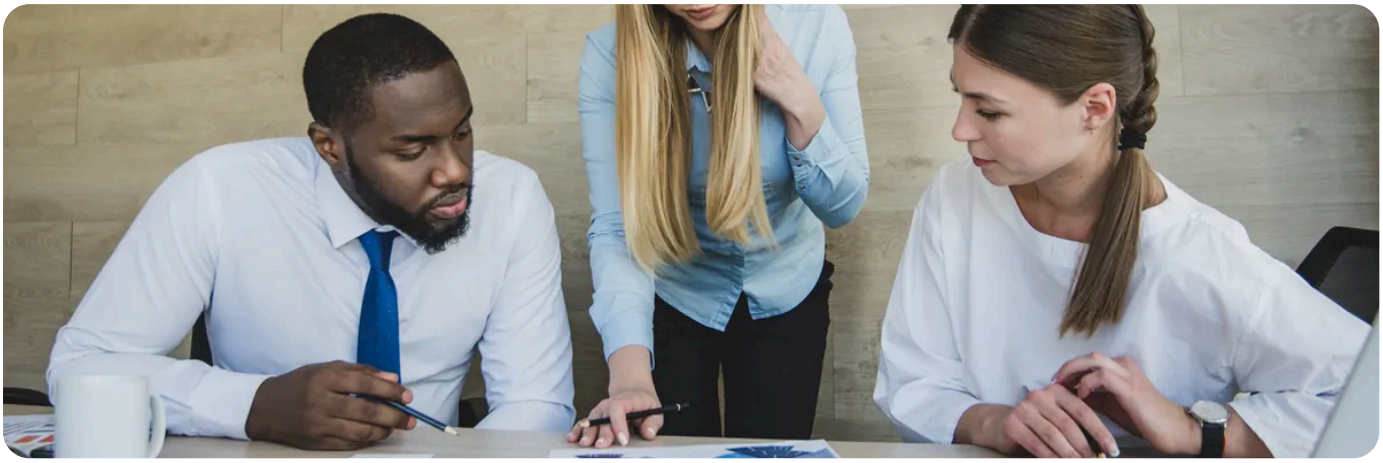
CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-9858	v2026.11.28	International · English

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/9858

The recommended coding models: the context and what each one is for

Choosing a Tool & Fixing Errors Created on **November 28, 2026** 3 min read



The All in One AI catalogue has many text models, but not all of them are good for coding. These six are in the catalogue, support tools (function calling) and are part of the Coding Plan. Use the id exactly as it is, with the slash:

Model	Context	What for
`moonshotai/kimi-k2.7-code`	262,144 tokens	Made for coding: the best starting point for everyday work.
`deepseek-ai/deepseek-v4-pro`	1,048,576 tokens	Strong reasoning, for large changes and hard bugs.
`zai-org/glm-5.2`	1,048,576 tokens	Built for agents: uses tools well in long tasks.
`qwen/qwen3.8-max`	1,000,000 tokens	A large model, for reasoning and code across whole projects.
`deepseek-ai/deepseek-v4-flash`	1,048,576 tokens	Fast and cheap, for background tasks and short questions.
`zai-org/glm-5.3-flash`	1,048,576 tokens	Light and cheap, good for simple edits.

Which one to choose

1. For everyday work, start with `moonshotai/kimi-k2.7-code`: it is made for coding.

2. For a large change or a hard bug, move to `deepseek-ai/deepseek-v4-pro`.

3. For long tasks, where the agent uses many tools in a row, use `zai-org/glm-5.2`.

4. To reason about a whole project, `qwen/qwen3.8-max`.

5. For background tasks and short questions, `deepseek-ai/deepseek-v4-flash`. It is the one our Claude Code setup puts in `ANTHROPIC_DEFAULT_HAIKU_MODEL`.

6. For simple, cheap edits, `zai-org/glm-5.3-flash`.

The context, and where to set it

The context (`context_length`) is the most tokens the model reads in one request, input and output together. `moonshotai/kimi-k2.7-code` reads 262,144 tokens; the other five, about a million. Give the tool the context of the model you chose, so it knows when the conversation reaches the limit:

Tool	Where you set the context
Claude Code	`CLAUDE_CODE_MAX_CONTEXT_TOKENS`
Cline	**Model Configuration**, with the maximum output
OpenCode and Kilo Code (file)	`limit.context` of each model
Aider	`max_input_tokens` in `.aider.model.metadata.json`

A bigger context does not make a request cheaper. Each request pays for the input tokens it carries: the longer the conversation, the more text goes in each request.

Output, price and the Coding Plan

The maximum output of each model is in `max_output_tokens` and the price in credits is in `pricing`. 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. With an active Coding Plan, these six spend points instead of credits, and stronger models spend more points per token. The models that spend plan points are listed in the **Plan tab of All in One AI**.

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.

Changing the model halfway

In Claude Code, type `/model zai-org/glm-5.2`. In Codex, `codex -m zai-org/glm-5.2` opens a session with another model. In OpenCode, use `/models`, in the provider/model format, for example `api/zai-org/glm-5.2`. If you move to `moonshotai/kimi-k2.7-code`, lower the context you set in the tool as well.

More models support tools: the model block in **the OpenCode article** has all 79, from the most expensive to the cheapest.

RECOMMENDED

Want to code with All in One AI models? Visit the Coding with AI page.

See Coding with AI →

SEE ALSO

- **Why a coding model must accept tools** →
- **Which coding tool to use** →
- **Choosing a text model** →
- **Coding Plan: which plan to choose** →
- **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-9858 · v2026.11.28 · Generated on 28/11/2026 21:53