Mozout

Generated on 28/11/2026 21:23

KNOWLEDGE BASE · PUBLIC DOCUMENT

aio chat options: --system, --max-tokens, stdin and the chat in the terminal

The aio Command

CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-9336	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/9336

aio chat options: --system, --max-tokens, stdin and the chat in the terminal

The aio Command Created on **November 28, 2026** 2 min read



`aio chat` asks a text model a question and shows the answer as the model writes it. The first steps are in **AI in the terminal with the aio command**. Here we go through the options, one by one.

Three ways to ask

```
aio chat "Explain what an MX DNS record is"
aio chat "Write a slogan for a bakery" --model deepseek-ai/deepseek-v4-flash --max-tokens 300
cat prompt.txt | aio chat
aio chat
```

Way	What happens
A prompt in quotes	The model answers and you get the terminal back.
Text through `|` (stdin)	Without a prompt in quotes, the text that comes from standard input is the prompt. Use it for a file or for the output of another command.
No prompt and no stdin	It opens a chat in the terminal. `/reset` starts over, and `/exit` or `Ctrl+C` quits.

The options

Option	What it does
`--model <id>`	Chooses the model. Without it, the `AIO_MODEL` variable applies, or the default model shown in `aio help chat`.

Option	What it does
`--system "..."`	Gives the model instructions, for example `--system "Answer in English"`.
`--max-tokens`	Limits the length of the answer.
`--no-stream`	Gets the whole answer at once, instead of bit by bit.
`--json`	The answer as JSON, to use in another program.

Everyday examples

```
cat error.log | aio chat --system "Explain this error in two sentences"
cat contract.txt | aio chat --system "Sum it up in five points" --max-tokens 400
```

`--system` sets the tone and the format. The file text arrives through `|` and is the prompt.

--max-tokens and your balance

Before calling the model, we reserve credits for the input and for the maximum output asked for. At the end, we charge the real tokens and return the rest. With a low balance, a lower `--max-tokens` helps, because the reservation is smaller. At the end of each answer, `aio` shows the input and output tokens.

With a Coding Plan, the plan models spend points instead of credits. The full picture is in **what a chat request costs through the API**.

With a `sandbox_` key, `aio chat` always answers "Resposta de teste (sandbox)." and spends nothing. Use it to test a script before you switch to the `prod_` key.

RECOMMENDED

Want to chat with AI models in the terminal? See All in One AI.

See All in One AI →

SEE ALSO

AI in the terminal with the aio command: installing, logging in and chatting →

What a chat request costs through the API: the reservation, max_tokens and the final adjustment →

- Seeing models and prices in the terminal: aio models and aio model
- Choosing a text model: price, context window and what each one is for
- All in One AI

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-9336 · v2026.11.28 · Generated on 28/11/2026 21:23