Mozout

Generated on 29/11/2026 04:04

KNOWLEDGE BASE · PUBLIC DOCUMENT

Context window and maximum output: what context_length and max_output_tokens mean

Playground & Text Models

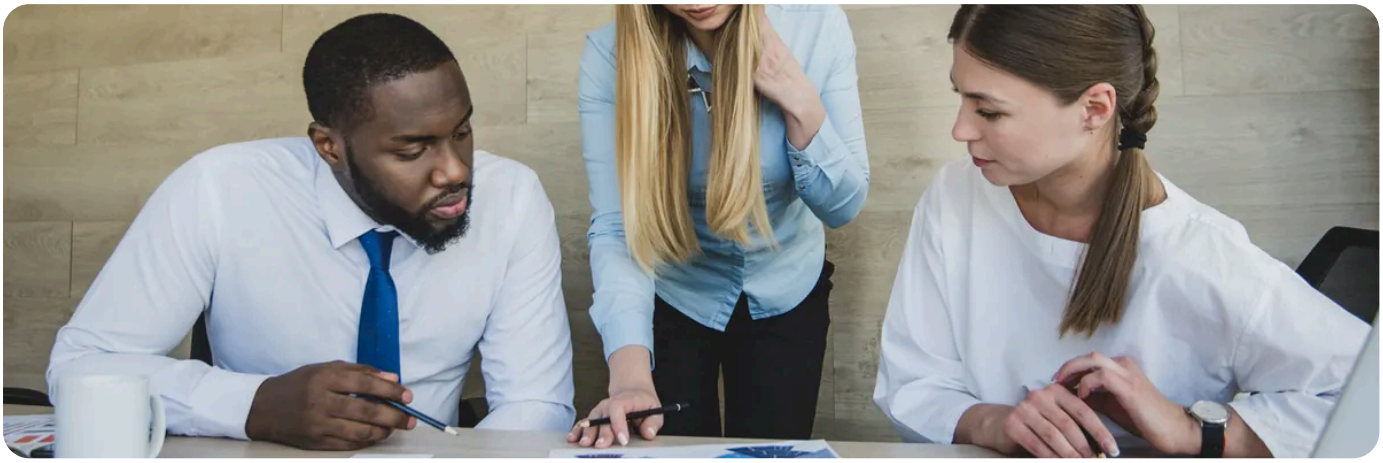
CODE	VERSION	COUNTRY · LANGUAGE
RLX-INT-KB-9042	v2026.11.28	International · English

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/9042

Context window and maximum output: what context_length and max_output_tokens mean

Playground & Text Models Created on **November 28, 2026** 2 min read



Every text model has two limits: how much it reads at once, `context_length`, and how much it writes in one answer, `max_output_tokens`. Both are counted in tokens.

The two numbers

Field	What it is
`context_length`	The most tokens the model reads in one request, input and output together.
`max_output_tokens`	The most tokens the model writes in one answer.

Two examples from the documentation: `deepseek-ai/deepseek-v4-flash` reads up to 1,048,576 tokens and writes up to 393,216; `moonshotai/kimi-k2.6` reads up to 262,144 and writes up to 262,144.

Where you see them

The **Models** tab shows the context of each text model ("Context: N tokens"). Both fields come in `GET /v1/models` and in the `aio model` command, followed by the id.

What counts towards the context

Everything in the request: the instructions, the whole conversation, the texts you pasted and the answer itself. In a conversation, each new message carries the earlier ones. For agents, the instructions and the knowledge go in every request. That is why a very long conversation ends up hitting the limit. The Playground shows "The conversation is too long. Please start a new one." See **why the AI forgets what you said earlier**.

max_output_tokens and your max_tokens

`max_output_tokens` is the model's ceiling. `max_tokens` is the limit you set on each request, below that ceiling. The credit reservation counts your `max_tokens`, or 4,096 if you do not send it. In the Playground, the "Maximum tokens in the reply" field takes 16 to 32,768, and the estimate never counts more than the model's maximum output.

In coding tools

When a tool asks for the context size, use `context_length`. For the maximum output, use a smaller value. 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.

RECOMMENDED

See the context of each model on the Models tab.

See All in One AI →

SEE ALSO

- Why the AI forgets what you said earlier: context and long conversations
- Choosing a text model: price, context window and what each one is for
- The Models tab: searching for a model and copying its id
- Tokens explained: how AI model APIs are counted
- 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-9042 · v2026.11.28 · Generated on 29/11/2026 04:04