MaaS

Model List

Issue 01

Date 2026-09-10

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 Model List

MaaS provides multiple models for you to use. You can easily integrate model services into your business by following tutorials or API documentation.

Model Recommendation

GLM-5.2	GLM-5.1
GLM-5.2 is Zhipu AI's next-generation flagship model, officially launched and open-sourced on June 17, 2026. Its core features include: Enhanced long-horizon task capability: reduces context drift and goal forgetting in complex tasks. State-of-the-art coding and long-horizon task evaluation: achieves open-source SOTA performance, delivering greater stability in complex system engineering and deep debugging. Improved real-world development experience: more reliable project-level context handling, adherence to engineering standards, and multi-platform development. Support for 1M sequence length, with maximum input of 1M, maximum output of 128K, and maximum CoT length of 64K. It supports chain-of-thought and function calling.	GLM-5.1 is the latest flagship model from Zhipu. It provides enhanced coding capabilities and improved performance in long-horizon tasks. It can autonomously work for up to 8 hours in a single task, covering the entire process from planning and execution to iterative optimization and delivering engineering-level results. While GLM-5.1 matches Claude Opus 4.6 in general intelligence and raw coding proficiency, it significantly outperforms the global frontier in long-horizon sustained execution. It excels at autonomous, multi-stage tasks over extended periods, making it the ideal foundation for building highly resilient autonomous agents and long-horizon coding engines. It supports sequence lengths up to 198K, with a maximum input of 192K, a maximum output of 128K, and a maximum thought chain of 96K. It supports chain-of-thought and function calling.

Text Generation

Tutorial: **Text Generation**; API: **Sending a Chat Request (Chat/Post)**; price: **MaaS Text Generation Models**.

Table 1-1 Model list

Model Parameter (Model ID)	Supported Capabilities	Length Limit	Default Rate Limit
glm-5.2	Deep thinking (configurable) Function calling	Context length: 1M Max input length: 1M Max output length: 128K Max chain-of-thought length: 64K	• TPM: 1,000,000 • RPM: 100
glm-5.1	Deep thinking (configurable) Function calling	Context length: 198K Max input length: 192K Max output length: 128K Max chain-of-thought length: 96K	• TPM: 1,000,000 • RPM: 100
deepseek-v4-flash	Deep thinking (configurable) Function calling Prefix completion	Context length: 1M Max input length: 1M Max output length: 384K Max chain-of-thought length: 96K	• TPM: 60,000 • RPM: 15
deepseek-v4-pro	Deep thinking (configurable) Function calling Prefix completion	Context length: 1M Max input length: 1M Max output length: 128K Max chain-of-thought length: 96K	• TPM: 30,000 • RPM: 3

Deep Thinking

Tutorial: **Deep Thinking**; API: **Sending a Chat Request (Chat/Post)**; price: **MaaS Text Generation Models**.

Table 1-2 Model list

Model Parameter (Model ID)	Supported Capabilities	Length Limit	Default Rate Limit
glm-5.2	Deep thinking (configurable) Function calling	Context length: 1M Max input length: 1M Max output length: 128K Max chain-of-thought length: 64K	• TPM: 1,000,000 • RPM: 100
glm-5.1	Deep thinking (configurable) Function calling	Context length: 198K Max input length: 192K Max output length: 128K Max chain-of-thought length: 96K	• TPM: 1,000,000 • RPM: 100
deepseek-v4-flash	Deep thinking (configurable) Function calling Prefix completion	Context length: 1M Max input length: 1M Max output length: 384K Max chain-of-thought length: 96K	• TPM: 60,000 • RPM: 15
deepseek-v4-pro	Deep thinking (configurable) Function calling Prefix completion	Context length: 1M Max input length: 1M Max output length: 128K Max chain-of-thought length: 96K	• TPM: 30,000 • RPM: 3

