

MaaS

Model List

Issue	01
Date	2026-09-21



Copyright © Huawei Cloud Computing Technologies Co., Ltd. 2026. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Cloud Computing Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI 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.....1

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.

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-5.3	Deep thinking (cannot be disabled) Function calling Prefix caching	Context length: 1M Max input length: 1M Max output length: 128K Max CoT length: 128K	- TPM: 1,000,000 - RPM: 100
glm-5.2	Deep thinking (configurable) Function calling	Context length: 1M Max input length: 1M Max output length: 128K Max CoT 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 CoT length: 96K	- TPM: 1,000,000 - RPM: 100

Model Parameter (Model ID)	Supported Capabilities	Length Limit	Default Rate Limit
deepseek-v4.1-flash	Deep thinking (configurable) Function calling Prefix caching	Context length: 1M Max input length: 1M Max output length: 384K Max CoT 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 CoT 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 CoT 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-5.3	Deep thinking (cannot be disabled) Function calling Prefix caching	Context length: 1M Max input length: 1M Max output length: 128K Max CoT length: 128K	• TPM: 1,000,000 • RPM: 100

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 CoT 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 CoT length: 96K	• TPM: 1,000,000 • RPM: 100
deepseek-v4.1-flash	Deep thinking (configurable) Function calling Prefix caching	Context length: 1M Max input length: 1M Max output length: 384K Max CoT 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 CoT 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 CoT length: 96K	• TPM: 30,000 • RPM: 3