

MaaS

Console Functions

Issue	01
Date	2026-10-08



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 Managing API Keys..... 1

2 Viewing a Built-in Model in Model Square..... 8

3 Using a Built-in Service..... 13

4 Creating an Endpoint..... 18

1 Managing API Keys

When using MaaS, you often need to call model services via APIs for tasks such as data requests and model inference. However, improper management of API keys (loss or leakage) can lead to unauthorized access risks, thereby compromising service security and stability. To prevent unauthorized access and minimize security risks, the MaaS platform provides comprehensive API key management. This feature allows you to create, edit, and control permissions, ensuring that only authorized users can call model services or manage keys. This section explains how to create, edit, and delete API keys.

Constraints

- Each account can have up to 30 API keys. If you exceed this limit, delete old keys before adding new ones.
- After you create an API key, you must save it right away. The system does not store the key as plain text. The key cannot be recovered if lost.
- After you delete an API key, it becomes invalid right away.

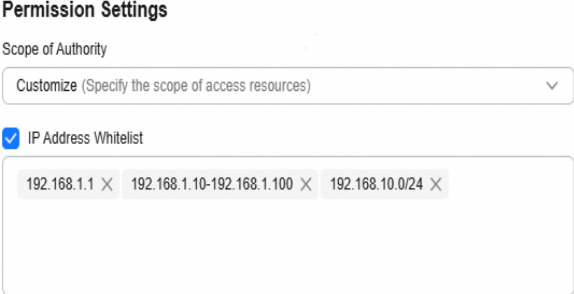
Creating an API Key

You can create up to 30 API keys. An API key is **displayed only once after being created. Copy and keep it secure.** If the key is lost, create a new API key.

1. Log in to the [MaaS console](#) and select the target region on the top navigation bar.
2. In the navigation pane, choose **Management and Statistics > API Key Management**.
3. On the **API Key Management** page, click **Create API Key**, configure the basic information and permission settings, and click **OK**.

The tag cannot be modified after the key is created.

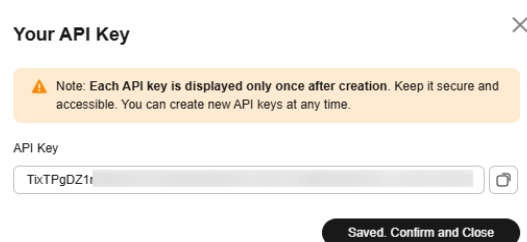
Table 1-1 Parameters

Parameter		Description
Basic information	Tag	Tag of the API key. The tag must be unique. The tag can contain 1 to 100 characters. Only letters, digits, underscores (_), and hyphens (-) are allowed.
	Description	Description of the custom API key. The value can contain 1 to 100 characters.
Permission Settings	Scope of Authority	You can select All (unrestricted access to all resources) or Customize (access to specified resources). The default value is All .
	IP Address Whitelist	This parameter can be configured when Scope of Authority is set to Customize. After this function is enabled, only IP addresses in the whitelist can be used to call model services. IPv4 CIDR blocks are supported. Enter an IPv4 CIDR block and press Enter. A maximum of 100 addresses can be entered. IPv4 example: ● 192.168.1.1 ● 192.168.1.10-192.168.1.100 ● 192.168.10.0/24 Figure 1-1 IP address whitelist Permission Settings  Only IP addresses in the whitelist can call model services. IP addresses you can still add: 97 Example: 192.168.1.1, 192.168.1.10-192.168.1.100, or 192.168.10.0/24.

Parameter		Description
	Model Service Access Scope	This parameter can be configured when Scope of Authority is set to Customize. You can select Built-in Services or Endpoint. • Built-in Services: The API key restricts access to selected models and versions when making calls using the model parameter. The system automatically includes all current and future built-in services and their versions. Choose specific ones if needed. • Endpoint: The system automatically includes all current and future endpoints. Choose specific ones if needed. For details about how to create an endpoint, see Creating an Endpoint.

4. In the displayed dialog box, read the message carefully, copy the key and save it to a secure location, and click **Saved. Confirm and Close**.

Figure 1-2 Saving an API key



After you close the dialog box, the key cannot be viewed again. Once you create the API key, it appears on the **API Key Management** page. The system shows only the first and last four characters by default.

Figure 1-3 API key



Editing an API Key

Once an API key is created, its tag cannot be modified. You can adjust other settings as needed. Modifying the permission may cause authentication failures for some model services and endpoints. Confirm the impact before performing the operation.

1. Log in to the [MaaS console](#) and select the target region on the top navigation bar.
2. In the navigation pane, choose **Management and Statistics > API Key Management**.

3. In the **Operation** column of the target API key, click **Edit**.
4. In the displayed dialog box, modify the basic information and permission settings as required, and click **OK**.

For details, see [Table 1-1](#).

5. In the **Edit API Key** dialog box, enter **YES** and click **OK**.

After an API key is updated, changes will take effect in 5 minutes.

Managing API Key Operation Permissions of IAM Users

If the administrator does not configure the API key operation permissions for IAM users, MaaS determines whether IAM users have such permissions based on strict authorization of the ModelArts console.

To view or modify strict authorization, follow these steps:

1. Log in to the [ModelArts console](#) as an administrator. In the navigation pane of the old console, choose **System Management > Permission Management**.

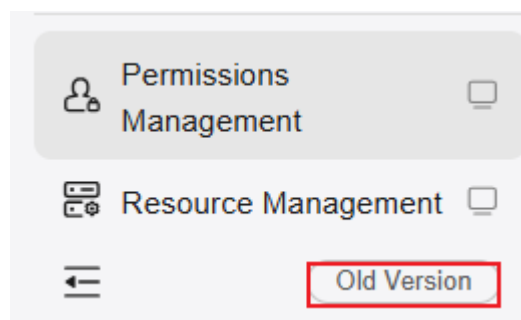
**System
Management**

Workspace

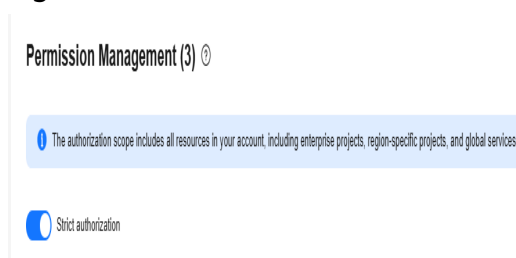
Permission
Management

NOTE

If you are using the new console, click **Old Version**.



2. Enable or disable strict authorization as required on the permissions management page.
 - Enabling strict authorization: IAM users who are not configured with the API key operation permissions do not have the API key management permissions. For details about how to configure the API key operation permissions for IAM users, see the following sections.
 - Disabling strict authorization: IAM users who are not configured with the API key operation permissions have the API key management permissions.

Figure 1-4 Strict authorization

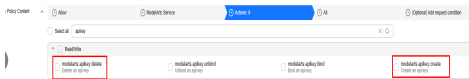
If the administrator configures the API key operation permissions for IAM users on the IAM console, MaaS determines whether IAM users have such permissions based on the administrator's permission configuration and strict authorization of the ModelArts console.

To configure the API key operation permissions for IAM users, follow these steps:

1. Create a custom policy.
 - a. Log in to the [IAM console](#) as an administrator. In the navigation pane, choose **Permissions > Policies/Roles**.
 - b. On the **Policies/Roles** page, click **Create Custom Policy** in the upper right corner.
 - c. On the **Create Custom Policy** page, configure parameters based on the table below and click **OK**.

Table 1-2 Parameters for creating a custom policy

Parameter	Description
Policy Name	Enter a policy name. Only letters, digits, spaces, and special characters (-_.,) are allowed.
Policy View	Click Visual editor .

Parameter	Description
Policy Content	Configure the API key permission for IAM users. - Select Allow or Deny. Allow: The user is allowed to perform operations on cloud services. Deny: The user is not allowed to perform operations on cloud services. - Select a cloud service, enter ModelArts in the search box, and select ModelArts Service (ModelArts). Figure 1-5 Selecting a cloud service  - Select an operation, enter apikey in the search box, and select the API key operation permissions for the IAM user as required. modelarts:apikey:delete: After this permission is granted, IAM users can delete API keys. modelarts:apikey:create: After this permission is granted, IAM users can create or edit API keys. Figure 1-6 Operation configuration 
Description	Enter a policy description.
Scope	Retain the default setting (project-level service).

2. Add the custom policy to the target user group.
- If you need to create a user group, see [Creating a User Group and Assigning Permissions](#).
- a. Log in to the [IAM console](#) as an administrator. In the navigation pane, choose **User Groups**.

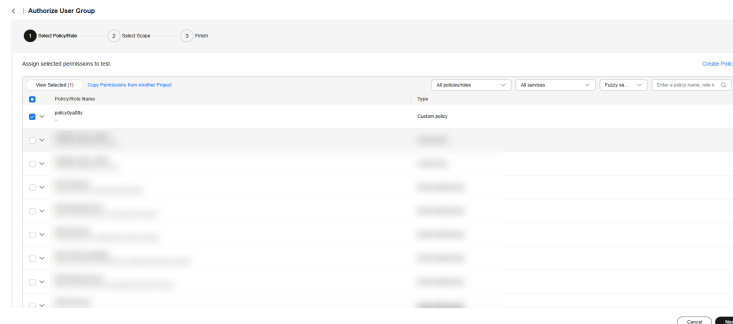
b. On the **User Groups** page, click **Authorize** in the **Operation** column of the target user group.

Figure 1-7 Authorize

Name	Users	Description	Created	Operation
test	2		Apr 08, 2025 16:20:17 GMT+08:00	Authorize Modify Manage User Delete

- c. On the displayed page, select the policy created in [step 1](#), click **Next**, select the authorization scope as required, and click **OK**.

Figure 1-8 Authorization page



3. Add the IAM user to the user group.
For details about how to create an IAM user, see [Creating an IAM User and Logging In](#).
 - a. Log in to the [IAM console](#) as an administrator. In the navigation pane, choose **User Groups**.
 - b. On the **User Groups** page, click **Manage User** in the **Operation** column of the user group modified in [step 2](#).
 - c. In the **Manage User** dialog box, select the target IAM user and click **OK**.
After an IAM user is granted required permissions, the IAM user can use the permissions selected in [step 1](#) on the **API Key Management** page of the MaaS console.

Deleting an API Key

If you reach the API key limit or if a key is leaked or lost, delete unused API keys. **Deleting the API key will permanently remove it and prevent its use or retrieval. Exercise caution when performing this operation.**

1. Log in to the [MaaS console](#) and select the target region on the top navigation bar.
2. In the navigation pane, choose **Management and Statistics > API Key Management**.
3. Click **Delete** on the right of the target API key.
4. In the **Delete API Key** dialog box, click **OK**.

FAQs

1. How long does it take for an API key to become valid after it is created?
An API key becomes valid a few minutes after creation.
2. Can I use an API key across regions?
API keys work only in their specific region. For example, an API key created in the CN-Hong Kong region can only be used in this region.

2 Viewing a Built-in Model in Model Square

MaaS provides various open-source models. You can check them on the **Model Square**. The model details page shows all necessary information. You can choose suitable models for training and inference to incorporate into your enterprise systems.

Prerequisites

- You have registered a Huawei Cloud account. For details, see [Signing Up for a Huawei Cloud Account](#).

Billing

Calling built-in services: Billing varies by model type. For details, see [Billing Items for MaaS Built-in Model Service](#).

Accessing the Model Square

- Log in to the [MaaS console](#) and select the target region on the top navigation bar.
- In the navigation pane, choose **Model Square**.
- In the **Filter** area on the **Model Square** page, filter models by type, context length, advanced features, series, and supported jobs, or search by model name.

For details about model series, see [Model Introduction](#). The supported filter criteria may vary depending on the region. The following uses CN-Hong Kong as an example.

Table 2-1 Model filters

Filter	Description
Type	Currently, only text models are available.

Filter	Description
Context Length	You can filter models by context length, including 256K–1M, 128K–256K, 23K–128K, and 0–32K. If you select multiple context lengths, the page displays the model set of the selected context lengths.
Advanced Capabilities	You can filter models based on their ability to handle function calls or deep reasoning.
Models	You can filter models by series, including DeepSeek, Qwen3, and GLM. If you select multiple models, the page displays the collection of the selected model series.
Source	You can filter models by their source.

4. Perform the following operations on the target model card on the **Model Square** page:
- [Viewing Built-in Model Details](#)

- [Calling APIs](#)

Model Introduction

The following table lists the models supported by the MaaS platform. For details about the models, go to the model details page.

Table 2-2 Models in the Model Square

Model Series		Type	Use Case	Supported Language	Model Introduction
GLM	GLM-5.3	Text generation	Reasoning, coding, and agentic tasks	Chinese and English	GLM-5.3 is the flagship foundation model released by Zhipu AI on August 14, 2026. It features 743 billion parameters, the MoE architecture, and a 1M-token context window. Built on the same foundation as GLM-5.2, its performance improvements come entirely from post-training scaling.

Model Series		Type	Use Case	Supported Language	Model Introduction
	GLM-5.2	Text generation	Reasoning, coding, and agentic tasks	Chinese and English	GLM-5.2 is the next-generation flagship model from Zhipu AI, open-sourced on June 17, 2026. With state-of-the-art open-source coding and long-horizon task capabilities, it significantly enhances engineering stability in real-world development and improves multi-platform delivery reliability.
	GLM-5.1	Text generation	Reasoning, coding, and agentic tasks	Chinese and English	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.
Deep Seek	DeepSeek-V4-Pro	Text generation	Q&A and text generation inference	Chinese and English	DeepSeek-V4-Pro is the flagship version of the DeepSeek-V4 series. It adopts a Mixture-of-Experts (MoE) architecture, with trillions of parameters and an ultra-long 1M (one million token) context window.

Model Series		Type	Use Case	Supported Language	Model Introduction
	DeepSeek-V4-Flash	Text generation	Q&A and text generation inference	Chinese and English	DeepSeek-V4-Flash is a lightweight, efficient version of the DeepSeek-V4 series. It supports a 1M-token context window while offering faster, cheaper API services through a smaller model size and fewer active parameters.
	DeepSeek-V4.1-Flash	Text generation	Q&A, text generation, and image understanding	Chinese and English	DeepSeek-V4.1-Flash is a Mixture-of-Experts (MoE) model with 552 billion backbone parameters and support for a 1M-token context window.
Qwen3	Qwen3	Text generation	Q&A	Chinese and English	The Qwen3 series includes LLMs and multimodal models created by the Qwen team. These models undergo extensive training using vast amounts of language and multimodal data, followed by fine-tuning with top-tier datasets.

Viewing Built-in Model Details

1. Log in to the [MaaS console](#) and select the target region on the top navigation bar.
2. In the navigation pane, choose **Model Square**.
3. In the **Filter** area on the **Model Square** page, filter models by type, context length, advanced features, or series, or search by model name.
4. Click a model card. On the **Model Details** page, view the model details, such as the model introduction, version information, and supported capabilities. In the upper right corner of the **Model Details** page, you can view the operations supported by the model.
 - [Calling APIs](#)

 NOTE

The operations supported by different models may vary. Check the console for details.

Calling APIs

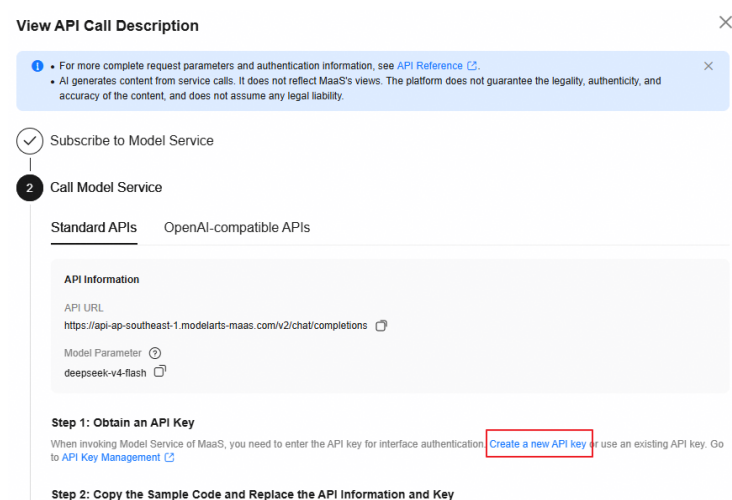
MaaS provides commercial-use APIs for inference. After subscribing to a model service, you can directly experience or call the model service without waiting for deployment. The model service is billed based on usage (such as the number of tokens or image generation duration).

1. Use either of the following methods to call an API:
 - On the **Model Square** page: Click **API Call** on a model card.
 - On the **Model Details** page: In the upper right corner of the page, click **API Calling** and select a version. Alternatively, click **API Calling** on the right of the version area.
2. On the **View API Call Description** page, enable or call the model service according to the instructions.

Subscribing to a model service: In the **Subscribe to Built-in Services and Try Popular Models Now** area, review the information, select **I have read and agree to the above terms and MaaS Service Statement**, and click **Subscribe**.

For details about the API key parameters, see [Creating an API Key](#). For details about the API calling parameters, see [Model API Calling Specifications](#).

Figure 2-1 View API Call Description



3 Using a Built-in Service

MaaS offers built-in services that give businesses high-performance, reliable inference APIs. Billing is based on token usage. Commercial services are suitable for scenarios that require high stability, frequent API calls, and professional support.

- Enterprise intelligent customer service: Enterprises hope to optimize their customer service systems by using inference APIs to achieve intelligent Q&A and intent recognition, thereby improving customer service efficiency and customer satisfaction.
- Content creation assistance: Media and advertising companies leverage inference APIs for copywriting and creative generation, enhancing the efficiency and quality of content production.
- Intelligent data analysis: Financial and e-commerce enterprises use inference APIs to conduct in-depth analysis of massive amounts of data, uncovering data value and aiding in decision-making.

Constraints

- This function is available only in the CN-Hong Kong region.
- Subscribing to a built-in service automatically subscribes you to all versions of that service. You cannot subscribe to just one version alone.
- Currently, built-in services cannot be disabled.

Billing

Model inference services process input by converting it into identifiable tokens. Using a built-in MaaS service charges you based on token count. For details, see [Billing Items for Model Inference](#).

Coupon Instructions

- When there is a discount offer, a relevant prompt will appear on the built-in service page. When subscribing to a built-in service, available coupons will be claimed by default. During the billing process, coupons will be applied first.
- The scope of application and eligibility criteria for different coupon activities vary. Whether you can successfully claim a coupon depends on the actual activity rules.

- For information on the issuance and usage of discount coupons for model services, go to **Billing Center > Promotions > Coupons**.

Calling Service APIs

Requests can be routed to different regions as needed.

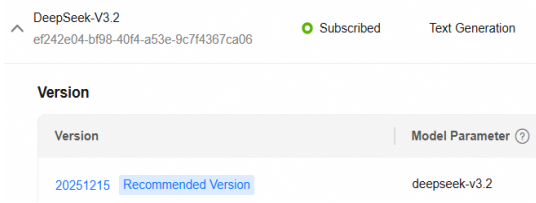
Subscribing to Built-in Services

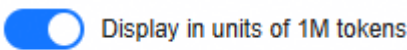
- Log in to the **MaaS console** and select the target region on the top navigation bar.
- In the navigation pane on the left, choose **Model Inference > Real-Time Inference**.
- Click the **Built-in Services** tab. In the **Operation** column of the target service, click **Subscribe**.
- Select the built-in service as required. Carefully read and agree to the terms, and then click **Subscribe**. The service's status in the built-in service list will change to **Subscribed**.

Selecting a built-in service will subscribe to all its model versions.

The following table describes the parameters in the built-in service list.

Table 3-1 Built-in service list parameters

Parameter	Description
Name/ Resource ID	Name and ID of the built-in service. - Check the service version and model parameter: Click  on the left of the service name to view the service version. The name displayed in the model Parameter column can be used for model calls. Figure 3-1 Viewing the model version  Click the version name to go to the Model Description page and view details about the model version. - View the service bill: Click  after the built-in service ID to copy the ID. On the Billing > Expenditure Details page of the Billing Center, view the bill details by resource ID.
Status	Subscribed: The built-in service is enabled. Not Subscribed: The built-in service is not enabled.
Type	Type of the built-in service

Parameter	Description
Billing Mode	Billing mode of the built-in service. The billing mode varies depending on the model. For details, see Billing Items .
Infer Price	Inference pricing of the built-in service. The inference pricing varies depending on the model. For details, see Billing Items. In the upper right corner of the page, enable or disable Million tokens unit display. Figure 3-2 Million tokens unit display  • If Million tokens unit display is disabled, the price is displayed by thousand tokens. • If Million tokens unit display is enabled, the price is displayed by million tokens.
Discounts	Discounts available for the built-in service. -- indicates that there is no discount.
Model flow limiting	Total quota for accessing all services under the same model within the current account. • TPM: The number of tokens processed per minute, including both input and output. • RPM: The number of requests processed per minute.
Call Statistics	Click  to go to the Service Call Details page and view the call data and monitoring metric details of the built-in service in a specified period. For more information, see Viewing the Call Data and Monitoring Metrics of Real-Time Inference on MaaS .

Parameter	Description
Operation	Operations supported by the built-in service • Unsubscribe: If this button is unavailable, the service cannot be disabled. No fee is generated when the service is not used. • API Calling: See Calling a Built-in Service. • ModelArts Guard: Configure content guard if you have a V2 or higher account level. After built-in services are enabled, ModelArts Guard is enabled by default. NOTE To view your account level, log in to the MaaS console, click your username in the upper-right corner, select Basic Information, and view the level displayed next to your account name. ModelArts Guard blocks harmful content in inputs and outputs during model inference. Enabling this feature may increase processing latency. When disabled, the models rely on their native security features. - To enable ModelArts Guard, click More and select Enable ModelArts Guard. In the displayed Enable ModelArts Guard dialog box, read the note and click OK. - To disable ModelArts Guard, click More and select Disable ModelArts Guard. In the displayed Disable ModelArts Guard dialog box, read the note, select I have read and agree to the above terms and Content Moderation Disclaimer, and click OK.

Calling a Built-in Service

1. Click **API Calling** in the **Operation** column of a subscribed built-in service and select a version.
2. On the **View API Call Description** page, call the built-in service as required.
 - For details about obtaining an API key, see [Managing API Keys](#).
 - API parameters vary depending on the model type. For details, see [MaaS API Call Specifications](#).
 - If an error occurs when you call a built-in service, locate the fault by referring to [Error Codes](#).

Traffic Control Rules

To ensure fairness in users' calls to the model, MaaS implements basic rate limiting. If a limit is exceeded, API requests will fail. You must wait until the rate limit resets before sending additional requests.

- **Tokens Per Minute (TPM):** The number of tokens (input + output) processed per minute.
- **Requests Per Minute (RPM):** The number of requests processed per minute.

It is recommended that you send API requests evenly to avoid sending a large number of requests in a short period. Rate limiting policies may be enforced based on QPS (RPM/60). Even if the total request volume is below the per-minute limit, sending a large burst of requests in a short period may trigger rate limiting. This will result in request failure and the return of error code ModelArts.81101 (Too Many Requests). For details about error codes, see [Error Codes](#).

FAQs

- **Is there any billing example?**

For details about billing items and examples, see [Prices of Real-Time Inference Model Services](#).

- **Can a built-in service be disabled after subscription?**

Built-in services cannot be disabled. However, no charges are incurred when a service is not used.

- **What should I do if the subscription status is frozen when I use a built-in service?**

For details about how to handle the frozen status, see [What Should I Do If the Subscription Status is Frozen When I Use a Built-in Service?](#)

- **What should I do if the subscription status is invalid when I use a built-in service?**

If the subscription status is invalid, the resource has been frozen due to overdue payments and the resource retention period has expired. You can subscribe to the built-in service again.

4 Creating an Endpoint

MaaS allows users to create endpoints. By invoking models through an endpoint name (in the **model** parameter), users can achieve traffic routing and fine-grained management across different business scenarios or model versions.

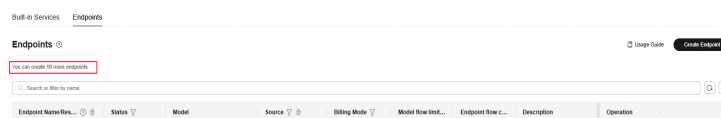
For LLM inference, the endpoint is the key access point for invoking a model. Developers can send API inference requests using a specified model ID, enabling fast access and flexible invocation.

During the development and operation of AI applications, enterprises and developers encounter issues such as chaotic inference service call management, difficulties in throttling, and ambiguous cost accounting. Multiple business lines might share the same inference service, leading to resource competition and unstable service performance. Additionally, the lack of effective call restrictions makes it challenging to track the resource consumption of each business line.

MaaS provides endpoints that let users create independent entry points, set rate-limiting rules, and track fees precisely by endpoint name. This helps manage inference resources effectively and optimize usage costs.

Constraints

- This function is only available in the CN-Hong Kong regions. Resources cannot be called across regions.
- This function allows you to create custom endpoints based exclusively on built-in services.
- The upper left corner of the **Endpoints** tab displays the number of remaining endpoints you can create.



- Each endpoint must have a unique name under an account. A deleted endpoint name cannot be used when you create a new endpoint.
- After an endpoint is created, the model service cannot be modified.
- The created endpoints must comply with the rules and specifications of the platform and cannot be called in violation of regulations.

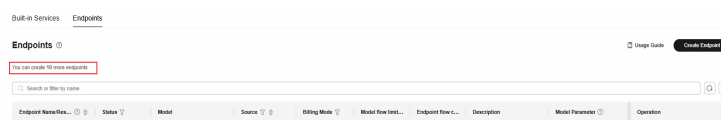
Billing

Endpoints are free to create. However, you may be charged for calling the model service or using resources. Check service costs in the Billing Center by endpoint name.

- Calling a custom MaaS endpoint: billed by MaaS. Token-based billing is supported. For details, see [Billing Items for MaaS Built-in Model Services](#).

Creating an Endpoint

1. Log in to the [MaaS console](#) and select the target region on the top navigation bar.
2. In the navigation pane on the left, choose **Model Inference** > **Real-Time Inference**. Click the **Endpoints** tab. You can view your remaining endpoint quota in the upper left corner.



3. Click the **Endpoint** tab and click **Create Endpoint** in the upper right corner of the page.
4. On the **Create Endpoint** dialog box, configure parameters.

Table 4-1 Parameters for creating an endpoint

Parameter		Description
Basic information	Parameter	Name of an endpoint. The endpoint name must be unique and cannot contain special characters. Enter 1 to 100 characters. The name must start with a digit or letter, and can only contain letters, digits, hyphens (-), underscores (_), and dots (.).
	Description	Description of an endpoint. The value can contain up to 256 characters.

Parameter		Description
	Please select a model	Select MaaS or ModelArts. MaaS: built-in models provided by MaaS. Click Please select a model. In the Select Model dialog box, select the model and version as required, and click OK. All built-in services, whether enabled or disabled, are supported by default. The supported access mode is displayed on the right of the model version. ModelArts: new-version real-time services are charged by ModelArts. Click Please select a model. In the Select Model Service dialog box, select the region, workspace, and service as required, and click OK. NOTE - You can only create endpoints for services with MaaS gateway enabled. The Select Model dialog box displays only services with MaaS gateway enabled. - Ensure that the selected service complies with OpenAI standards and specifications. Otherwise, the call may fail.
Billing Information	Billing Mode	The following billing modes are available:
Rate Limiting	Model flow limiting	After you select a built-in model, the total traffic limit for the model under the current account is displayed. Requests Per Minute (RPM): The number of requests processed per minute. Tokens Per Minute (TPM): The number of input and output tokens processed per minute.

Parameter		Description
	Endpoint Flow Control	You can manually set the traffic limit for an endpoint. If the combined rate limit of all endpoints accessing the same model under an account equals the model's total quota, it effectively prevents competition for traffic allocation among different endpoints. - You can configure different RPM and TPM traffic controls for each endpoint, as long as they stay within your account's traffic limits. - The values of RPM and TPM must be positive integers.

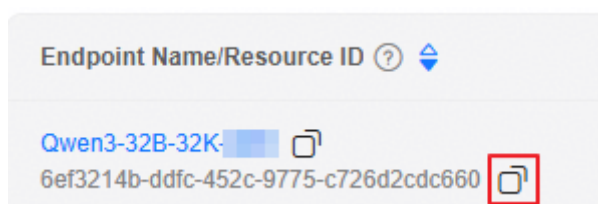
- On the right of the page, view the configuration summary, select **I have read and agree to MaaS Service Statement**, confirm the billing information at the bottom of the page, and click **Create Now**.
- Once you create the endpoint, its details appear in the **Endpoint** tab. From there, you can see that its status is **In use** and call or experience it online.

Figure 4-1 Endpoint created

Endpoint Name/Resource ID	Status	Model	Source	Billing Mode	Model Flow Limit	Endpoint Flow Limit	Description	Operation
Qwen3-32B-32K	In use	MaaS	...	Token-based billing	TPM: 999,999,999 RPM: 999,999,999	TPM: 20 RPM: 5	...	API Calling Try More
...	In use	MaaS	...	Token-based billing	TPM: 999,999,999 RPM: 999,999,999	TPM: 20 RPM: 5	...	API Calling Try More

To check bill details by ID, click  next to the endpoint resource ID to copy the ID. Go to the **Billing Center**, choose **Billing > Transactions and Detailed Bills**, and click **Bill Details**.

Figure 4-2 Copying an ID



Viewing Endpoint Details

- On the **MaaS console**, choose **Model Inference > Real-Time Inference**. On the displayed page, click the **Endpoints** tab. In the **Endpoint Name/Resource ID** column of the target endpoint, click the endpoint name or resource ID. The **Overview** page is displayed by default. You can view details about the endpoint, including basic information, billing information, rate limiting information, and more configurations. For details about the parameters, see [Table 4-2](#).

Table 4-2 Parameters about an endpoint

Parameter		Description
Basic information	Endpoint Name	Name of an endpoint.
	ID	ID of an endpoint.
	Model	Model selected for the endpoint. Click a model name to view the model information.
	Source	Source of the model selected for the endpoint. Currently, MaaS and ModelArts are supported.
	Model Parameter	Model parameter selected for the endpoint.
	Created At	Time when an endpoint was created.
	Description	Description of an endpoint.
Billing Information	Billing Mode	Billing mode of the endpoint. It is token-based billing. For details, see Table 4-1 .
	Inference Price	Inference pricing unit. You can switch between 1M tokens and 1K tokens.
Flow Control	Model Flow Limiting	This parameter is displayed only when Select Model is set to MaaS . After you select a built-in model, the total traffic limit for the model service under the current account is displayed. • Requests Per Minute (RPM): The number of requests processed per minute. • Tokens Per Minute (TPM): The number of input and output tokens processed per minute.

Parameter		Description
	Endpoint Flow Control	You can manually set the traffic limit for an endpoint. If the combined rate limit of all endpoints accessing the same model under an account equals the model's total quota, it effectively prevents competition for traffic allocation among different endpoints. - You can configure different RPM and TPM traffic controls for each endpoint, as long as they stay within your account's traffic limits. - The values of RPM and TPM must be positive integers.
More Configurations	ModelArts Guard	Displays the current status of the content safety guardrails. It is On or Off .
	Implicit Watermark	This parameter is displayed only when Select Model is set to MaaS and a built-in video or image generation service is selected. This function is disabled by default. Once enabled, you must add metadata-based implicit identifiers to AI-generated content.
	Explicit Watermark	This parameter is displayed only when Select Model is set to MaaS and a built-in video or image generation service is selected. This function is disabled by default. Once enabled, model outputs will include an AI watermark.

- Endpoints can be edited, deleted, enabled, disabled, and called via API.
 - To check bill details by ID, click  next to the endpoint resource ID to copy the ID. Go to the [Billing Center](#), choose **Billing > Transactions and Detailed Bills**, and click **Bill Details**.
 - You can also click the  icon next to the model parameter to copy the model parameter. On the API Calling tab page, you can call different access points based on different model parameters.
2. Click the **View API Call Description** tab to view the process of calling an endpoint. For details, see [Calling an Endpoint](#).
 3. Click the **Monitoring** tab to view the call details of the endpoint.
 - On the **Monitoring** tab, select the time range, time granularity, calling method, and IP address as required.

Table 4-3 Filters for endpoint monitoring

Parameter	Description
Time Range	Service call statistics can be viewed for today, yesterday, the past 3, 7, or 14 days, or across a custom date range. Time range and precision filtering rules: • For time ranges of 1–2 days, metrics can be queried at minute or hour precision. • For time ranges of 3–7 days, metrics can be queried at hour or day precision. • For time ranges longer than 7 days, metrics can be queried at day precision only.
Time granularity	The calling statistics can be collected by minute or hour.
Calling Method	API key calls and online experience are supported. • By default, all API keys are used to authenticate calls to a model service on MaaS. You can choose specific ones if needed.
IP Address	The client source IP address (public IP) that has generated calls is derived from the http_x_forwarded_for field in the APIG logs. When this field contains multiple values, the system will use the first value; if the field value is -, it will be displayed as an empty string. The IP address is displayed by default as All, but you can also select specific IP addresses as needed.

- Usage: Track endpoints' token usage to prevent overuse.

Table 4-4 Usage metrics

Parameter	Description
Completed Tasks	Number of completed tasks at the endpoint.
Total Inferences	Number of inferences at the endpoint.
Total Requests	The total number of requests at the endpoint.
Total Tokens Used (K Tokens)	The total number of tokens used by the endpoint within each unit of time.
TPM (K Tokens/Minute)	The number of input and output tokens processed per minute.
RPM	The number of requests processed per minute.

Parameter	Description
QPS	The number of queries processed per second. When Time Precision is set to By the minute , the value displayed is the peak QPS recorded during that minute.
Input Tokens (K Tokens)	The length of input tokens. • AVG: The average length of input tokens. • MAX: The maximum length of input tokens. • P50: 50% of the input tokens are below this value. • P80: 80% of the input tokens are below this value. • P90: 90% of the input tokens are below this value. • P99: 99% of the input tokens are below this value.
Output Tokens (K Tokens)	The length of output tokens. • AVG: The average length of output tokens. • MAX: The maximum length of output tokens. • P50: 50% of the output tokens are below this value. • P80: 80% of the output tokens are below this value. • P90: 90% of the output tokens are below this value. • P99: 99% of the output tokens are below this value.

- Performance metrics: View performance metrics to identify opportunities for improvement.

Table 4-5 Performance metrics

Parameter	Description
Average Generation Time (s)	The average time taken to generate each image/video.

Parameter	Description
Average Response Latency (ms)	The end-to-end latency of successful requests per unit time • AVG: The average end-to-end latency. • MAX: The maximum end-to-end latency. • P50: 50% of the end-to-end latencies are below this value. • P80: 80% of the end-to-end latencies are below this value. • P90: 90% of the end-to-end latencies are below this value. • P99: 99% of the end-to-end latencies are below this value.
Time to First Token (ms)	The time from receiving a request to generating the first output token. The statistics are collected only for streaming responses. Due to model version constraints, certain versions cannot display these metrics in non-streaming scenarios. Upgrade to the latest version to view these metrics. • AVG: The average TTFT. • MAX: The maximum TTFT. • P50: 50% of the TTFTs are below this value. • P80: 80% of the TTFTs are below this value. • P90: 90% of the TTFTs are below this value. • P99: 99% of the TTFTs are below this value.

Parameter	Description
TPOT (ms)	The interval between two output tokens. The statistics are collected only for streaming responses. Due to model version constraints, certain versions cannot display these metrics in non-streaming scenarios. Upgrade to the latest version to view these metrics. AVG: The average TPOT. • MAX: The maximum TPOT. • P50: 50% of the TPOTs are below this value. • P80: 80% of the TPOTs are below this value. • P90: 90% of the TPOTs are below this value. • P99: 99% of the TPOTs are below this value.
Average Task Processing Duration (min)	Average time spent on processing tasks.

- Cache: View the cache hit metrics of an endpoint to evaluate the system performance.

Table 4-6 Cache hit metrics

Parameter	Description
Cache Hits (K Tokens)	The total number of tokens served from the cache at the endpoint.
Cache Hit Rate (%)	The proportion of input tokens served from the endpoint cache, expressed as a percentage of total input tokens.

- Troubleshooting: Quickly identify issues such as sudden increases in calls, abnormal consumption, and call failures during specific time periods.

Table 4-7 Abnormal metrics

Parameter	Description
Request Failure Rate (%)	The proportion of failed requests to the total number of requests.
Total Failed Requests	Indicates the number of failed requests. An error message is displayed when a request fails.

Parameter	Description
Inference Failure Rate (%)	The proportion of failed inferences to the total number of inferences.

- For more information, see [Viewing the Call Data and Monitoring Metrics of Real-Time Inference on MaaS](#).

Calling an Endpoint

An endpoint can be called only when it is in the **In use** status. The content generated by AI through service calls does not reflect the views of MaaS. The platform does not guarantee the legality, authenticity, or accuracy of the content and assumes no legal liability.

1. On the [MaaS console](#), choose **Model Inference > Real-Time Inference** and click the **Endpoints** tab. In the **Operation** column of the target endpoint, click **API Calling**.
2. On the **View API Call Description** page, obtain the API key as prompted, copy the call example, replace the API information and API key, and call the API.
 - The **model Parameter** column in the **Endpoint** tab shows the model's name used in the code during service calls. You can call different endpoints based on different **model** parameters.
 - For details about how to create an API key, see [Creating an API Key](#).
 - For details about the parameters in the call example, see [MaaS Standard API V2](#).

Managing an Endpoint

You can view call data and metric data, and edit, enable, or disable endpoints.

Viewing the Call Statistics of an Endpoint

View call data and metrics for a custom endpoint during a chosen time frame. This includes call counts, failures, and token totals. These insights help learn about service usage, monitor performance shifts, assess models, identify issues, fix problems, and improve efficiency.

1. On the [MaaS console](#), choose **Model Inference > Real-Time Inference** and click the **Endpoints** tab.
2. Click  in the **Call Statistics** column of the target endpoint to go to the **Service Call Details** page and check the details.

For more information, see [Viewing the Call Data and Monitoring Metrics of Real-Time Inference on MaaS](#).

Editing an Endpoint

You can modify the information about an endpoint, such as the description and traffic limit. **The model service of an endpoint cannot be modified.**

1. On the [MaaS console](#), choose **Model Inference > Real-Time Inference** and click the **Endpoints** tab. In the **Operation** column of the target endpoint, choose **More > Edit**.
2. In the **Edit Endpoint** dialog box, modify parameters as required and click **Update**.

For details about the parameters, see [Parameters for creating an endpoint](#).

Disabling or Enabling an Endpoint

An endpoint can be disabled when its status is **In use**. When you disable an endpoint, its inference feature turns off but can be reactivated later. Because of billing delays, you might continue receiving bills for the service even after disabling it.

An endpoint can be enabled when its status is **Disabled**. Once enabled, invoking or experiencing the endpoint will incur charges based on the original billing mode.

1. On the [MaaS console](#), choose **Model Inference > Real-Time Inference**, click the **Endpoints** tab, and perform the following operations as required.
 - Disabling an endpoint
 - i. Choose **More > Disable** in the **Operation** column of the target endpoint.
 - ii. In the displayed dialog box, click **OK**.
After the endpoint is disabled, the status of the endpoint is **Disabled**.
 - Enabling an Endpoint
 - i. Choose **More > Enable** in the **Operation** column of the target endpoint.
 - ii. In the displayed dialog box, click **OK**.
After the endpoint is enabled, its status is **In use**.

Deleting an Endpoint

You can delete an endpoint if it is no longer needed. **After an endpoint is deleted, its inference is disabled, and all data is deleted and cannot be restored. Exercise caution when performing this operation.**

Because of billing delays, you might continue receiving bills for the service even after deleting it.

1. On the [MaaS console](#), choose **Model Inference > Real-Time Inference** and click the **Endpoints** tab. In the **Operation** column of the target endpoint, choose **More > Delete**.
2. In the displayed dialog box, confirm the information, enter **DELETE**, and click **OK**.

After the endpoint is deleted, it is no longer displayed on the **Endpoints** page.

FAQs

- **What should I do if the number of endpoints I have created reaches the limit?**

You can delete the endpoints that are no longer needed and then create new ones.

- **How can I determine the number of tokens consumed?**

On the **Call Statistics** page, check the total number of tokens used, number of input tokens, and number of output tokens. For details, see [Viewing the Call Statistics of an Endpoint](#).

- **How long does it take for changes to the traffic limiting settings of an endpoint to take effect?**

After you save the changes, the traffic limiting settings will take effect immediately, and subsequent calls will be executed according to the new rules.

- **What should I do if the billing status of my endpoint changes to invalid?**

If the billing status of your endpoint becomes invalid, it indicates that the resource was frozen due to account arrears and exceeded the resource retention period, resulting in deletion. You may proceed to delete this endpoint and recreate it as needed.