PanguLM Service

User Guide

Issue 01

Date 2025-07-28



HUAWEI

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Copyright © Huawei Cloud Computing Technologies Co., Ltd. 2025. 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


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 Process of Using PanguLM..... 1

2 Preparations.....4

2.1 Applying for Trial Use of ModelArts Studio Large Model Development Platform.....4

2.2 Subscribing to the PanguLM Service.....4

2.3 Configuring Service Access Authorization.....10

2.4 Creating and Managing Workspaces.....11

2.4.1 Workspace Overview.....11

2.4.2 Creating and Managing Workspaces.....11

2.4.3 Managing Workspace Members.....14

3 Accessing Models in the Model Square..... 20

4 Using Data Engineering to Create a Dataset..... 23

4.1 Introduction to Data Engineering.....23

4.2 Process of Using Data Engineering.....28

4.3 Dataset Format Requirements.....29

4.3.1 Format Requirements for Text Datasets.....29

4.3.2 Format Requirements for Image Datasets.....33

4.3.3 Format Requirements for Video Datasets.....51

4.3.4 Format Requirements for Audio Datasets.....54

4.3.5 Format Requirements for Weather Datasets.....55

4.3.6 Format Requirements for Other Datasets.....58

4.4 Importing Data to the Pangu Platform.....58

4.5 Processing Datasets.....61

4.5.1 Dataset Processing Scenarios.....61

4.5.2 Processing Text Datasets.....65

4.5.2.1 Processing Text Datasets.....65

4.5.2.2 Synthesizing Text Datasets.....69

4.5.2.3 Labeling Text Datasets.....79

4.5.2.4 Combining Text Datasets Based on a Specific Ratio.....85

4.5.3 Processing Image Datasets.....89

4.5.3.1 Processing Image Datasets.....89

4.5.3.2 Labeling Image Datasets.....92

4.5.3.3 Combining Image Datasets Based on a Specific Ratio.....98

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.5.4 Processing Video Datasets 100
4.5.4.1 Processing Video Datasets 100
4.5.4.2 Labeling Video Datasets 102
4.5.5 Processing Weather Datasets 107
4.5.6 Processing Audio Datasets 110
4.5.6.1 Processing Audio Datasets 110
4.5.6.2 Labeling Audio Datasets 111
4.5.7 Processing Other Datasets 115
4.5.8 Managing Processing Operators 116
4.5.8.1 Introduction to Preset Processing Operators 116
4.5.8.1.1 Text Dataset Processing Operators 116
4.5.8.1.2 Video Dataset Processing Operators 134
4.5.8.1.3 Image Dataset Processing Operators 145
4.5.8.1.4 Weather Dataset Processing Operators 151
4.5.8.1.5 Audio Dataset Processing Operators 152
4.5.8.2 Custom Data Processing Operators 157
4.5.9 Managing Processed Datasets 171
4.5.10 Managing Processing Task Resources 172
4.5.11 Generating a Dataset in a Processing Task 172
4.6 Publishing a Dataset 173
4.6.1 Dataset Publishing Scenarios 173
4.6.2 Publishing Text Datasets 175
4.6.2.1 Evaluating Text Datasets 175
4.6.2.2 Publishing Text Datasets 177
4.6.3 Publishing Image Datasets 179
4.6.3.1 Evaluating Image Datasets 179
4.6.3.2 Publishing Image Datasets 182
4.6.4 Publishing Video Datasets 184
4.6.4.1 Evaluating Video Datasets 184
4.6.4.2 Publishing Video Datasets 187
4.6.5 Publishing Audio Datasets 188
4.6.6 Publishing Weather Datasets 189
4.6.7 Publishing Other Datasets 190
4.6.8 Managing Published Datasets 192
4.7 Converting the Dataset Format 193
4.7.1 Constraints 193
4.7.2 Format Conversion Process 194
4.7.3 Conversion Operators 196
4.8 Common Errors and Solutions for Data Engineering 201

5 Developing a DeepSeek Model 204

5.1 DeepSeek Models 204
5.2 Using Data Engineering to Build a DeepSeek Model Dataset 205

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

5.3 Training a DeepSeek Model 206
5.3.1 Creating a DeepSeek Model Training Job 207
5.3.2 Viewing the Training Status and Metrics of the DeepSeek Model 219
5.3.3 Publishing the Trained DeepSeek Model 221
5.3.4 Managing DeepSeek Model Training Jobs 221
5.4 Compressing a DeepSeek Model 222
5.5 Deploying a DeepSeek Model 223
5.6 Evaluating a DeepSeek Model 225
5.6.1 Creating a DeepSeek Model Evaluation Dataset 225
5.6.2 (Optional) Creating an API 228
5.6.3 Creating a DeepSeek Model Evaluation Job 233
5.6.4 Viewing the DeepSeek Model Evaluation Report 241
5.6.5 Managing DeepSeek Model Evaluation Jobs 243
5.7 Calling a DeepSeek Model 244
6 Developing a Prompt Engineering Project 247
6.1 What Is Prompt Engineering? 247
6.2 Obtaining a Prompt Template 249
6.3 Writing Prompts 249
6.3.1 Creating a Prompt Engineering Project 249
6.3.2 Writing Prompts 250
6.3.3 Previewing Prompt Outcomes 252
6.4 Comparing Outcomes Between Prompts 253
6.4.1 Setting Candidate Prompts 253
6.4.2 Comparing Outcomes Between Prompts 254
6.5 Evaluating the Prompt Outcomes in Batches 255
6.5.1 Creating a Prompt Evaluation Dataset 255
6.5.2 Creating a Prompt Evaluation Task 257
6.5.3 Viewing the Prompt Evaluation Result 258
6.6 Publishing Prompts 259
7 Developing an Agent 260
7.1 Agent Development Platform Overview 260
7.1.1 Platform Overview 260
7.1.2 Application Scenarios 262
7.1.3 Function Overview 264
7.1.4 Procedure 267
7.2 Using a Preset Agent in the Application Library 268
7.3 Agent Setup and Configuration 269
7.3.1 Quickly Setting Up an Agent 269
7.3.2 Improving Skills for Agents 276
7.3.2.1 Configuring a Plug-in 276
7.3.2.2 Configuring a Workflow 278
7.3.2.3 Configure a Knowledge Base 280

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

7.3.2.4 Configuring an MCP Service 282
7.3.2.5 Improving Dialog Experience of Agents 283
7.3.2.6 Debugging an Agent 285
7.3.3 Using APIs to Call an Agent 288
7.3.4 Managing Agents 290
7.4 Workflow Development and Configuration 293
7.4.1 Workflow Introduction 293
7.4.2 Creating a Workflow 295
7.4.2.1 Creating a Workflow 295
7.4.2.2 Debugging and Publishing a Workflow 297
7.4.2.3 Dialogue-based Workflows and Task-based Workflows 298
7.4.2.4 Workflow Node Configuration Reference 301
7.4.2.4.1 Start and End Nodes 301
7.4.2.4.2 LLM Node 304
7.4.2.4.3 Knowledge Repo Node 306
7.4.2.4.4 IntentDetection Node 308
7.4.2.4.5 Plugin Node 315
7.4.2.4.6 Branch Node 321
7.4.2.4.7 Code Node 324
7.4.2.4.8 Message Node 327
7.4.2.4.9 Questioner Node 329
7.4.2.4.10 Loop Node 335
7.4.2.4.11 Variable Assignment Node 338
7.4.2.4.12 Aggregation Node 340
7.4.2.4.13 Input Node 343
7.4.2.4.14 Workflow Node 345
7.4.2.4.15 MCP Service Node 346
7.4.2.4.16 Agent Node 348
7.4.3 Using APIs to Call a Workflow 351
7.4.4 Managing Workflows 354
7.5 Plug-in Development and Configuration 356
7.5.1 Introduction to Plug-ins 356
7.5.2 Creating a Plug-in 357
7.5.2.1 Creating a Plug-in Based on an API 357
7.5.2.2 Importing Plug-ins Using JSON Files 360
7.5.3 Managing Plug-ins 360
7.6 Knowledge Base Development and Configuration 361
7.6.1 Knowledge Base Introduction 361
7.6.2 Creating a Knowledge Base and Uploading Documents 364
7.6.3 Knowledge Base Hit Test 366
7.6.4 Managing Knowledge Bases 366
7.7 MCP Service Development and Configuration 367

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

7.7.1 MCP Service Introduction 367
7.7.2 Creating an MCP Service 367
7.7.3 Managing MCP Services 374
7.7.4 Subscribing to the MCP Service 375
7.8 Common Errors and Solutions During Agent Development 376
8 Managing Workspace Assets 382
8.1 Introduction to Pangu Model Workspace Assets 382
8.2 Managing Pangu Data Assets 382
8.3 Managing Pangu Model Assets 386
8.4 Managing Pangu Component Assets 387
9 Managing Resource Pools 388
9.1 Creating an Edge Resource Pool 388

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

1 Process of Using PanguLM

Introduction to the PanguLM Service

The PanguLM service includes Pangu foundation models, Pangu industry-specific model, and ModelArts Studio large model development platform. This "model + model development platform" product portfolio enables customers across various industries to rapidly develop and deploy large models in a one-stop manner.

ModelArts Studio Large Model Development Platform is a comprehensive large-model development platform introduced by the PanguLM service. It seamlessly integrates data management, model training, and model deployment, enabling developers to efficiently develop and deploy large models with ease. This platform is equipped with three toolchains: data engineering, model development, and application development. These toolchains empower developers to fully leverage the capabilities of Pangu models.

- **Data engineering toolchain**: Data is the cornerstone of large model training. As an integral part of the platform, the data engineering toolchain provides functionalities like data acquisition, data processing, and data publishing, helping ensure the high quality and consistency of data. This toolchain efficiently collects and processes data in various formats to meet diverse training job requirements. Additionally, it offers robust data storage and management capabilities, providing a solid foundation for large model training.
- **Model development toolchain**: This toolchain is a core component of the PanguLM service, offering a comprehensive solution from model creation to deployment. It includes model training, compression, deployment, evaluation, and calling, ensuring an efficient workflow for model application.
- **Application development toolchain**: This toolchain is an important module of the Pangu model platform. It supports prompt engineering and agent development, significantly speeding up the development of large model applications. This enables enterprises to swiftly address complex service requirements.

Process of Using Data Engineering

ModelArts Studio provides the data engineering capability to help you construct high-quality datasets, facilitating better model predictions and decision-making.

Figure 1-1 and **Table 1-1** illustrate the process of using data engineering.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 1-1 Flowchart of using data engineering

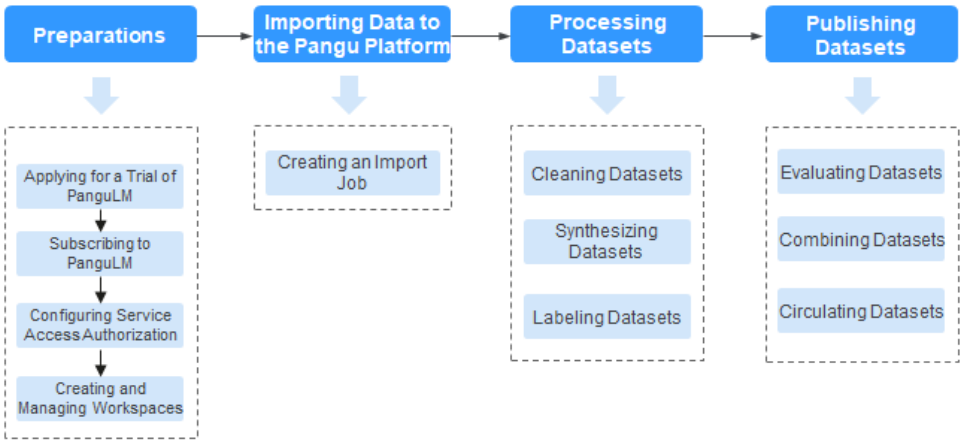


Table 1-1 Process of using data engineering

Procedure	Step	Description
Preparations	Applying for a trial of the PanguLM service	The PanguLM service is available for trial use. You can use this service only after your trial application is approved.
	Subscribing to the PanguLM service	Before using the PanguLM service, you need to subscribe to the service.
	Configuring service access authorization	To support data storage and model training, you need to grant PanguLM the permission to access OBS.
	Creating and managing Pangu workspaces	Create and centrally manage workspaces.
Importing data to the Pangu platform	Creating an import task	Import data stored in OBS into the platform for centralized management, facilitating subsequent processing or publishing.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Procedure	Step	Description
Processing datasets	Processing datasets	Use dedicated processing operators to preprocess data, ensuring it meets the model training standards and service requirements. Different types of datasets utilize operators specially designed for removing noise and redundant information, to enhance data quality.
	Generating synthetic datasets	Using either a preset or custom data instruction, process the original data, and generate new data based on a specified number of epochs. This process can extend the dataset to some extent and enhance the diversity and generalization capability of the trained model.
	Labeling datasets	Add accurate labels to unlabeled datasets to ensure high-quality data required for model training. The platform supports both manual annotation and AI pre-annotation. You can choose an appropriate annotation method based on your needs. The quality of data labeling directly impacts the training effectiveness and accuracy of the model.
	Combining datasets based on a specific ratio	Dataset combination involves combining multiple datasets based on a specific ratio and generating a processed dataset. A proper ratio ensures the diversity, balance, and representativeness of datasets and avoids issues resulting from uneven data distribution.
Publishing datasets	Evaluating datasets	The platform offers predefined evaluation standards for multiple types of data. You can choose from these predefined standards or customize evaluation standards as needed to precisely improve data quality, ensure that data meets high standards, and enhance model performance.
	Publishing datasets	Data publishing refers to publishing a single dataset in a specific format as a published dataset for subsequent model training operations. Datasets can be published in standard or Pangu format. ● Standard format: default dataset format on the platform. Datasets in this format can be published as assets, but cannot be used in the development of Pangu models. ● Pangu format: When training a Pangu model, you need to publish it in this format. Currently, only text and image datasets can be published in Pangu format.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2 Preparations

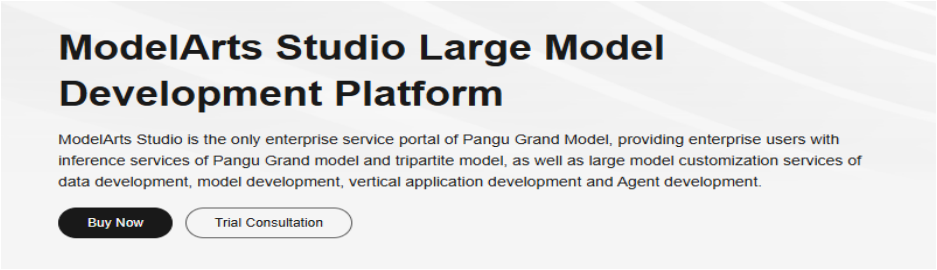
2.1 Applying for Trial Use of ModelArts Studio Large Model Development Platform

ModelArts Studio is available for trial.

To apply for trial use, perform the following steps:

1. Log in to ModelArts Studio Large Model Development Platform.
2. Click **Trial Consultation** to go to the Huawei Cloud presales support page.

Figure 2-1 Applying for trial



3. Enter user information such as the name and phone number, and click **Submit**.

2.2 Subscribing to the PanguLM Service

Description

ModelArts Studio allows you to subscribe to **model assets, data resources, training resources, and inference resources** on a yearly/monthly basis. You can also subscribe to resources on a yearly/monthly or pay-per-use basis.

- Model assets: **Model assets** can be used in the model development and application development modules. Currently, NLP models can be subscribed to.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Data resources: **Data general computing units** are used for regular expression operator processing during data processing. **Data intelligent computing units** are used for AI operator processing during data processing. **Data hosting units** are used for storing datasets during data engineering.
- Training resources: **Training units** can be used for model training and compression of all large models.
- Inference resources: **Inference units** can be used for model inference of NLP and CV models. **Model instances** can be used for model inference of scientific computing models.

Subscription Scenario

ModelArts Studio provides **free models and paid models**. After subscribing to the models and resources, you can use the **data engineering**, **model development**, and **agent development** functions on the platform.

You can subscribe to products on demand based on application scenarios.

Table 2-1 Subscription scenarios

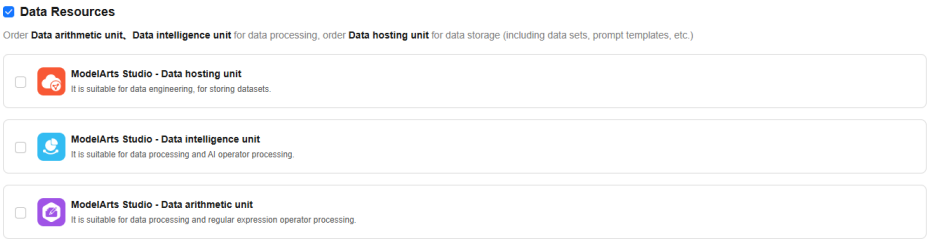
No.	Scenario	Description
1	Data scenario	Only data resources are subscribed to. Models are not subscribed to. Scenarios such as data processing, synthesis, labeling, and combination are supported.
2	Inference scenario	In the pure inference scenario, you can use free models or subscribe to model assets (optional) and inference resources for inference after deployment.
3	Training and inference scenario	Model training and inference are supported. You can subscribe to model assets (optional), training resources, and inference resources. Subscribed paid or free models can be used for training, and the trained models can then be deployed for inference to achieve better performance.
4	Agent scenario	Subscribe to model assets (optional), training resources (optional), and inference resources. You can use the paid or free models that have been subscribed to in the agent.

Scenario 1: Data scenario (Use only the data engineering capability.)

Subscribe to data general computing unit and data intelligent computing units for data processing, synthesis, labeling, and ratio-based storage. Subscribe to data hosting units for data storage (including datasets and prompt templates).

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

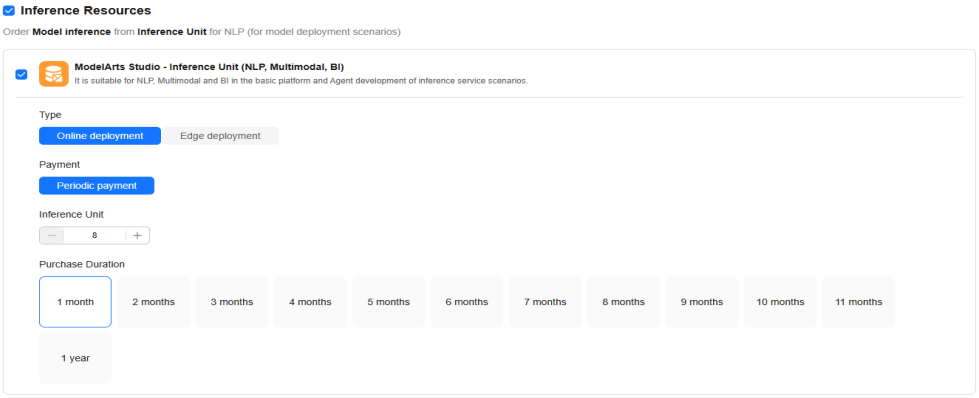
Figure 2-2 Data resources



Scenario 2: Inference scenario (Deploy and perform inference using a subscribed paid model or a free model in Model Square.)

Subscribe to ModelArts Studio - Inference Unit (NLP) for the deployment and inference of free and paid NLP models.

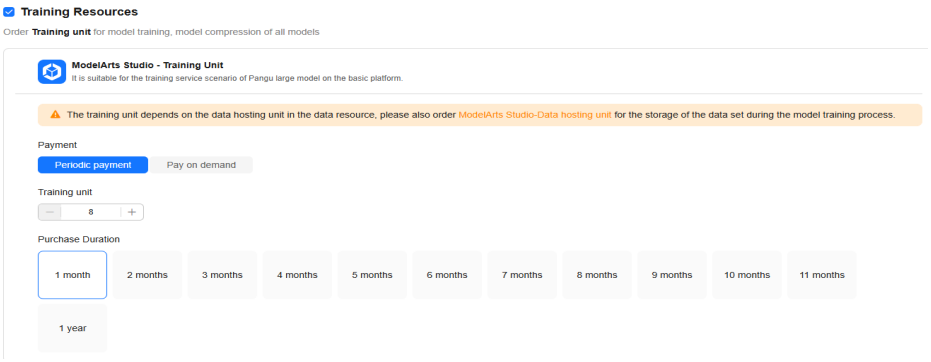
Figure 2-3 Inference scenario



Scenario 3: Training and inference scenario (Subscribe to model assets, training resources, and inference resources to train and deploy models.)

Subscribe to ModelArts Studio - Training Unit, use the subscribed paid model or a free model in Model Square to train the model, and deploy the model after the training is complete.

Figure 2-4 Training resources

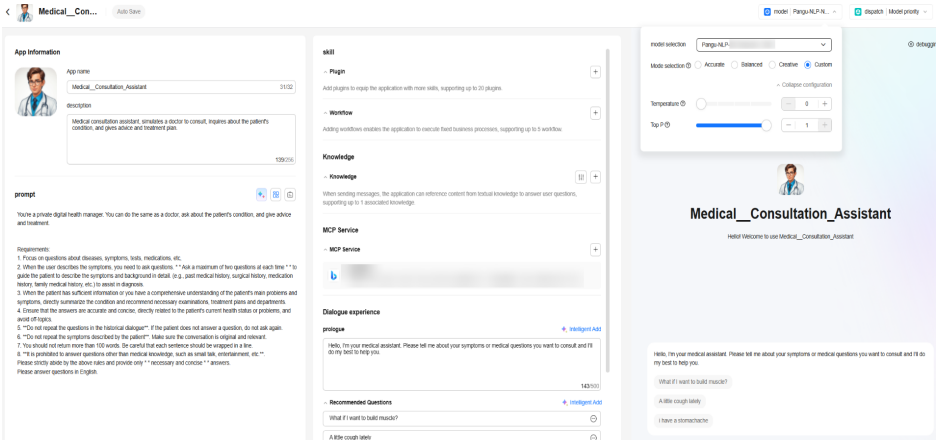


Scenario 4: Agent scenario (Use a deployed model for an agent.)

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Subscribe to model assets (optional), training resources (optional), and inference resources. You can use the paid or free models that have been subscribed to in the agent.

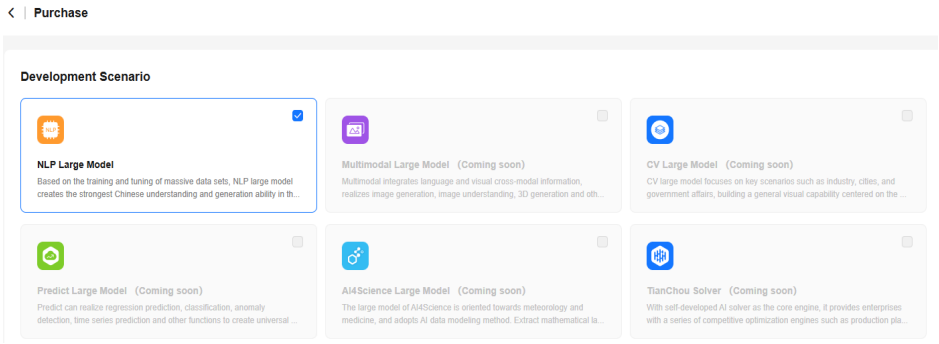
Figure 2-5 Agent



Procedure

1. Log in to ModelArts Studio Large Model Development Platform using a master account and click **Order Now**.
2. In the **Development Scenario** area, select one or more large models.

Figure 2-6 Selecting a development scenario



3. Subscribe to data resources, training resources, and inference resources by referring to **Table 2-2**.

Table 2-2 Resource subscription

Resource Name	Subscription Item	Application Scenario	Billing Mode
Data resources	ModelArts Studio - Data Hosting Unit	Stores data (including datasets and prompt templates).	Yearly/Monthly

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Resource Name	Subscription Item	Application Scenario	Billing Mode
	ModelArts Studio - Data General Computing Unit	Applies to regular expression operator processing during data processing. For details about the data resource types required by different data processing operators, see **Introduction to Preset Processing Operators**.	Pay-per-use (duration) or yearly/monthly
	ModelArts Studio - Data Intelligent Computing Unit	Applies to AI operator processing during data processing. For details about the data resource types required by different data processing operators, see **Introduction to Preset Processing Operators**.	Pay-per-use (duration) or yearly/monthly
Training resources	ModelArts Studio - Training Unit	Applies to model training and compression of all models.	Pay-per-use (duration) or yearly/monthly
Inference resources	ModelArts Studio - Inference Unit (NLP)	Applies to the inference service scenario of NLP models on the basic platform and agent development.	Yearly/Monthly

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Renewing Model Assets

ModelArts Studio supports the renewal of model assets in **yearly/monthly** mode, allowing you to extend the usage period of your subscribed model assets.

To renew a model asset, perform the following steps:

1. Log in to ModelArts Studio and click **Order Management** in the upper right corner of the page.
2. On the **Order Management** page, click the **Model ordering** tab. In the subscription model list, click **Renew** in the **Operation** column.
3. On the **Renewal Management** page, renew the model asset as prompted.

Unsubscribing from a Model Asset

To unsubscribe from a model asset, perform the following steps:

1. Log in to ModelArts Studio and click **Order Management** in the upper right corner of the page.
2. On the **Order Management** page, click the **Model ordering** tab. In the subscription model list, click **Unsubscribe** in the **Operation** column.

NOTE

Unsubscribing from a model asset does not affect running tasks such as model training, compression, evaluation, and deployment. However, after the model asset is unsubscribed, the model cannot be selected to create a task. Exercise caution when performing this operation.

Unsubscription is a high-risk operation. Before unsubscribing from a model asset, ensure that you have saved all necessary data and progress to avoid unnecessary losses.

Purchasing Additional Model Assets

In ModelArts Studio, you can purchase additional model assets based on your current subscription. Once the purchase is completed, you can utilize multiple model assets.

To purchase additional model assets, perform the following steps:

1. Log in to ModelArts Studio and click **Order Management** in the upper right corner of the page.
2. On the **Order Management** page, click **New purchase** n the upper right corner.
3. On the **Purchase** page, purchase new model assets by referring to **Procedure**.

Renewing Resources

ModelArts Studio supports **yearly/monthly** subscriptions for data, training, and inference resources.

To renew resources, perform the following steps:

1. Log in to ModelArts Studio and click **Order Management** in the upper right corner of the page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. On the **Order Management** page, click the **Resource ordering** tab. In the resource list, click **Renew** in the **Operation** column.
3. On the **Renewal Management** page, renew the resource as prompted.

NOTE

The pay-per-use billing mode does not support renewal.

Unsubscribing from Resources

To unsubscribe from a resource, perform the following steps:

1. Log in to ModelArts Studio and click **Order Management** in the upper right corner of the page.
2. On the **Order Management** page, click the **Resource ordering** tab. In the resource list, click **Unsubscribe** in the **Operation** column.

NOTE

Unsubscription will stop running services, and the resource will be unavailable.

Unsubscription is a high-risk operation. Before unsubscribing from resources, ensure that you have saved all necessary data and progress to avoid unnecessary losses.

Scaling Resources

ModelArts Studio supports scaling of data, training, and inference resources.

To scale resources, perform the following steps:

1. Log in to ModelArts Studio and click **Order Management** in the upper right corner of the page.
2. On the **Order Management** page, click the **Resource ordering** tab. In the resource list, click **Scaleup** in the **Operation** column.
3. On the **Scaleup** page, complete the scaling of the current resource. The platform will pay or refund the price difference based on the specifications before and after the scaling.

NOTE

Scale-in may affect the creation of ongoing tasks and subsequent tasks. Before scale-in, ensure that the resources to be scaled in have been released.

2.3 Configuring Service Access Authorization

Configuring OBS Access Authorization

ModelArts Studio uses OBS to store data. It provides secure, reliable, and cost-effective storage. To store data and train models, you need to configure the permission to access OBS.

To configure the OBS access permission, perform the following steps:

1. Log in to ModelArts Studio.
2. Configure OBS access authorization.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Method 1: Click **here** in the pop-up message on the top of the homepage. In the dialog box that is displayed, select the authorization item and click **Confirm**.

  Figure 2-7 Configuring OBS access authorization

- Method 2: Click **Setting** in the upper right corner of the homepage. On the **Authorization management** tab page, click **One-click authorization**.

2.4 Creating and Managing Workspaces

2.4.1 Workspace Overview

Workspaces enable flexible, efficient asset management and collaboration. The platform allows you to create independent workspaces based on your business needs or team structures.

Each workspace is completely isolated at the asset level to ensure asset security and operation independence, effectively avoiding risks caused by cross interference or incorrect permission configuration. You can create multiple workspaces based on actual scenarios, such as project management, department operation, and specific R&D requirements, to implement refined asset management and orderly allocation, helping you efficiently plan and assign tasks and improving team collaboration efficiency.

In addition, the platform is equipped with a comprehensive role permission system, covering multiple roles such as super administrators, administrators, and model development engineers. With flexible permission settings, each user can securely and efficiently operate platform functions within the corresponding permission scope, maximizing data security and work efficiency.

2.4.2 Creating and Managing Workspaces

Constraints

If the user group to which the current user belongs does not have the **Tenant Administrator (administrator of all cloud services)** permission, the workspace cannot be created.

To grant the **Tenant Administrator** permission to the current user before creating a workspace, you can either assign the Tenant Administrator permission to the

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

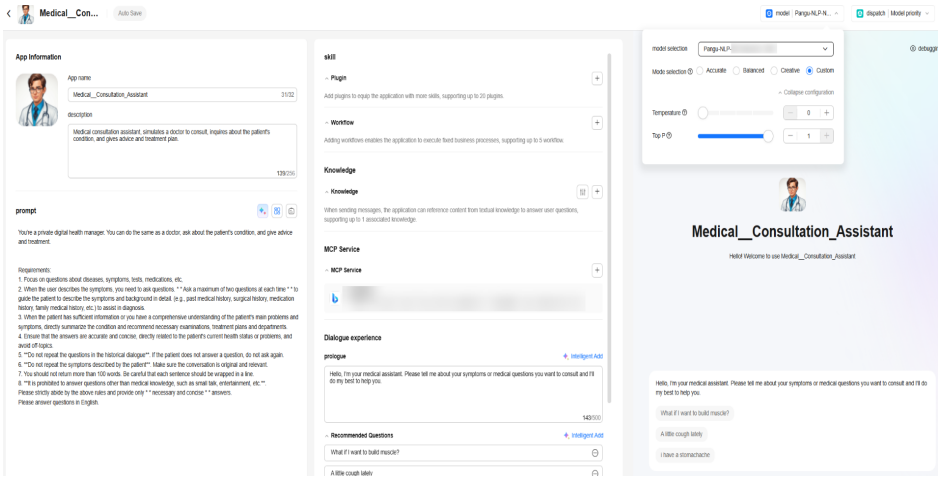
current user group by referring to **Creating a User Group** and **Creating a Pangu User**, or create a user group with this permission and add the user to that group.

Creating a Workspace

The process for creating a workspace is as follows:

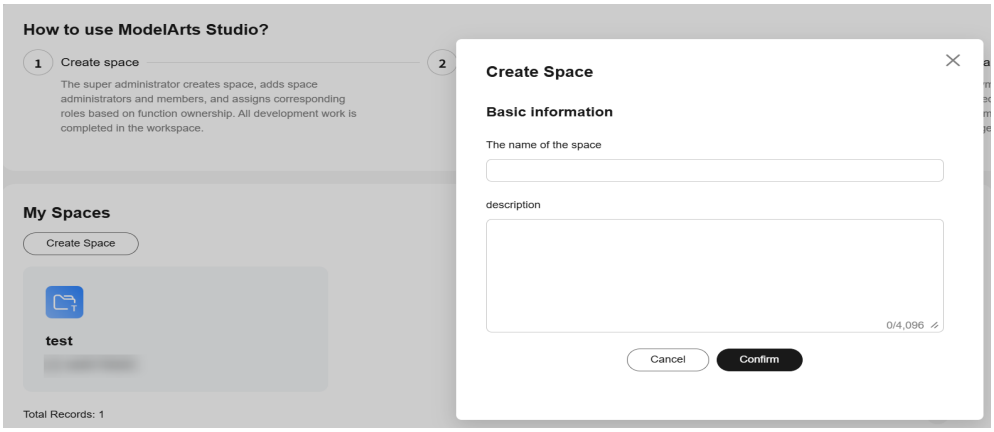
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click **Create space**.

   Figure 2-8 My Spaces

2. Enter the workspace name and description, and click **Confirm**.

   Figure 2-9 Creating a workspace

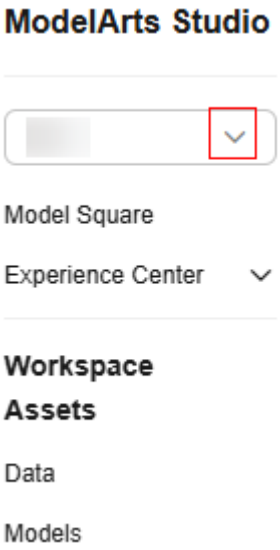
   

3. Click the created workspace to access ModelArts Studio.

   If you have the permission to access multiple workspaces, you can click in the upper left corner of the page to switch workspaces.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 2-10 Switching the workspace

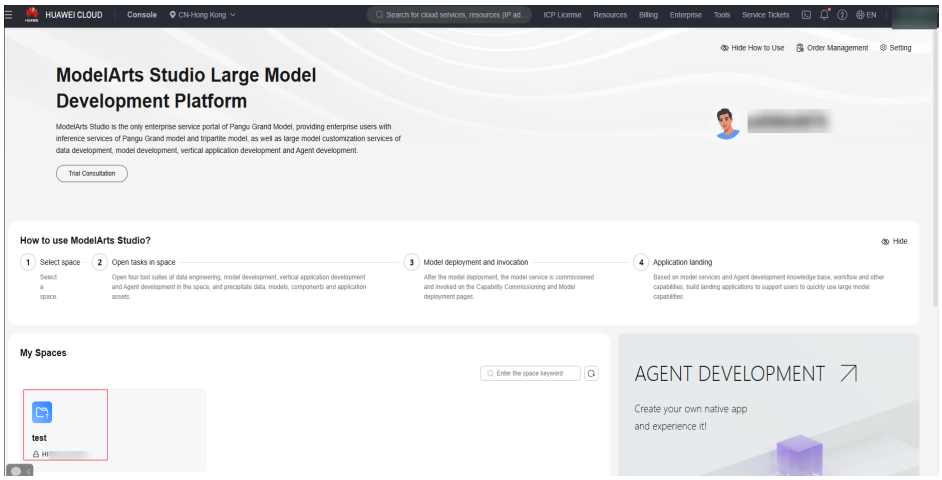


Managing Pangu Workspaces

You can view details about the current Pangu workspace, modify the workspace name and description, and delete unnecessary workspaces.

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 2-11 My Spaces



2. In the navigation pane on the left, click **Workspace Management**. On the **Workspace Settings** tab page, perform the following operations:
   - Modify the name or description of the current workspace.
   - View the creation time of the current workspace.
   - Click **Delete** in the upper right corner to delete the current workspace.

   NOTE

   Before deleting a workspace, ensure that the workspace is no longer used.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2.4.3 Managing Workspace Members

To isolate function usage permissions from asset permissions and ensure asset security and management efficiency, you can configure different roles for employees, setting distinct access permissions according to their responsibilities.

If your Huawei Cloud account does not require individual IAM users for permissions management, skip this section.

You can use IAM and the member management function of ModelArts Studio to implement fine-grained permissions management for IAM users.

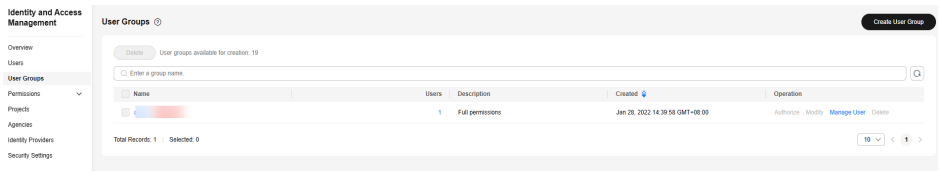
Creating a User Group

As an administrator, you can create user groups and grant them permissions using policies or roles. Users added to the user groups inherit permissions from the user groups.

To create a user group, do as follows:

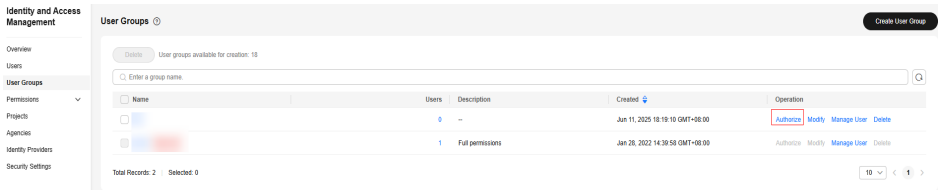
1. Log in to the **IAM console** using a master account.
2. On the IAM console, choose **User Groups** from the navigation pane, and click **Create User Group** in the upper right corner.

Figure 2-12 Creating a user group



3. On the **Create User Group** page, enter a user group name, and click **OK**.
4. Return to the user group list and click **Authorize** in the **Operation** column.

Figure 2-13 Authorizing a user group



5. Search for the required actions using the search box, select the actions, and click **Next**. For details, see **Table 2-3**.

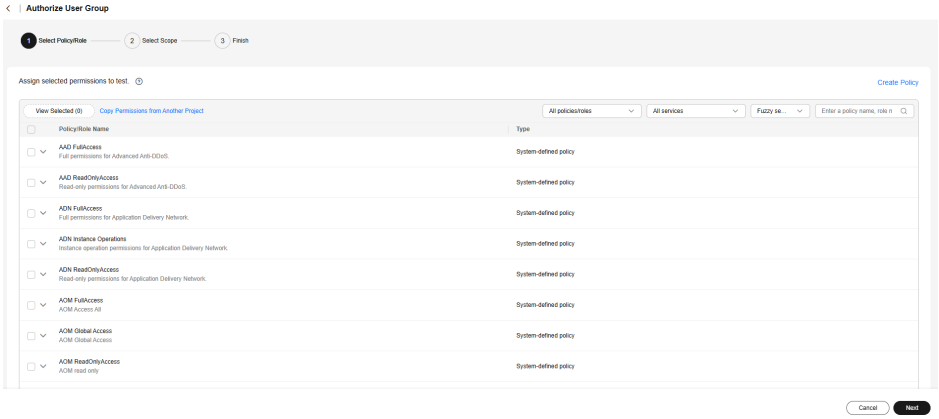
Table 2-3 Permissions

Permissions	Description
Agent Operator	Users with this permission can switch their roles to the delegating account to access the authorized services.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Permissions	Description
Tenant Administrator	Administrator permissions for all cloud services (except IAM)
Security Administrator	Full permissions on IAM (except for switching roles).

Figure 2-14 Adding permissions for the user group



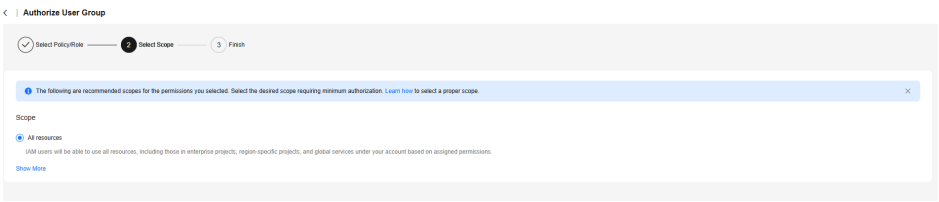
6. Select a scope.

   The system automatically recommends an authorization scope based on the policy of the selected action.

   - **All resources**: IAM users in the user group can use all enterprise projects, region-specific projects, and global services in the account based on the permissions assigned.
   - **Region-specific projects**: Select a region, for example, CN Southwest-Guiyang1. Then, IAM users in the user group can only use resources in the region-specific projects.
   - **Global services**: Global services are deployed in all regions, so you can access them seamlessly without having to switch between regions. However, global services cannot be authorized based on regional projects. They include services like Object Storage Service (OBS) and Content Delivery Network (CDN).

   Click **OK**.

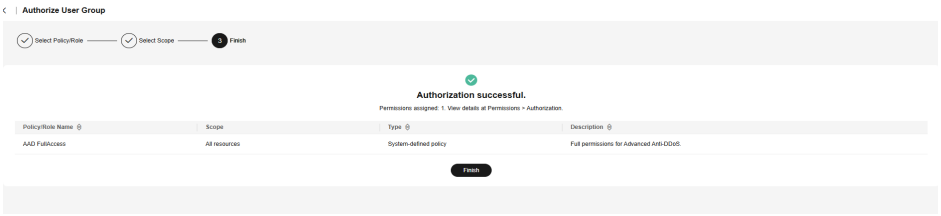
Figure 2-15 Selecting a scope



7. Click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 2-16 Authorization completed

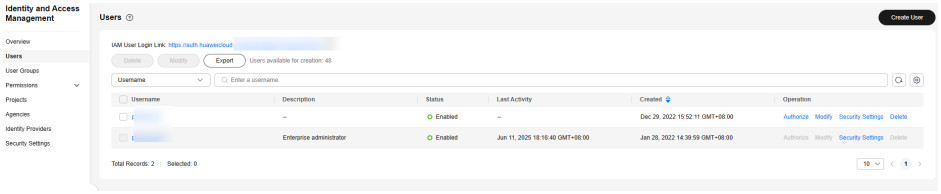


Creating a Pangu User

To create a Pangu user, do as follows:

1. Log in to the **IAM console** using a master account.
2. Choose **Users** from the navigation pane, and click **Create User** in the upper right corner.

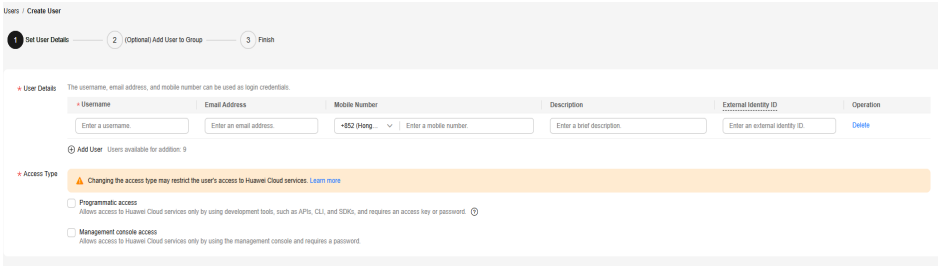
Figure 2-17 Creating a user



3. Configure basic information and click **Next**.

   When configuring user information, select **Programmatic access** for **Access Types**. If this option is not selected, IAM users cannot use PanguLM service APIs and SDKs.

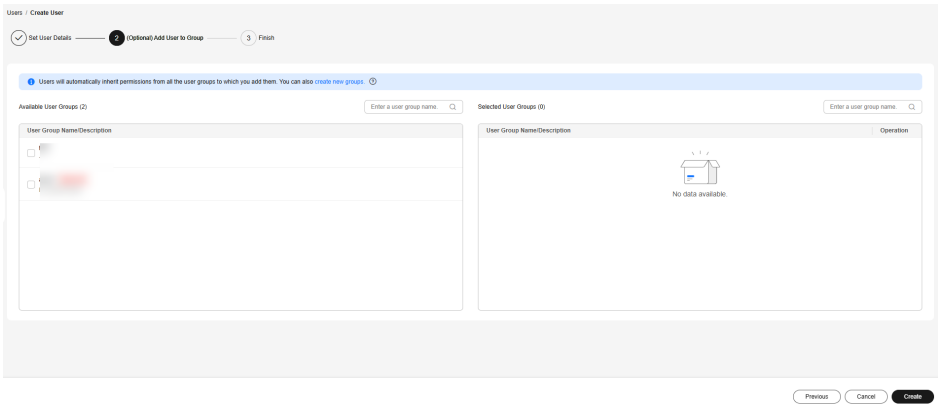
Figure 2-18 Configuring basic information about the user



4. Add the user to the user group created in **Creating a User Group** and click **Create**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 2-19 Adding the user to a user group

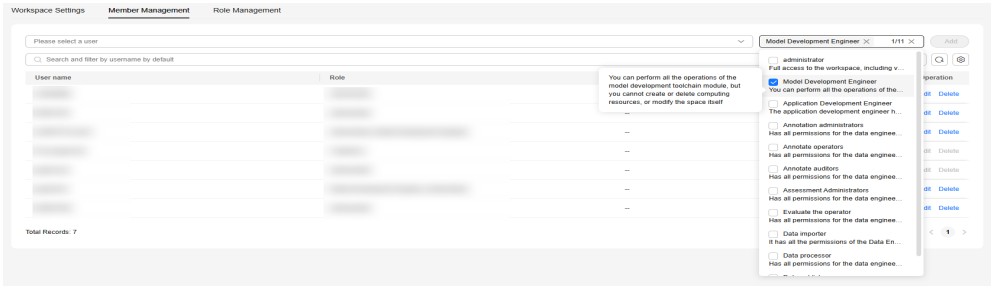


Adding a Pangu User to a Workspace

You can add a created Pangu user to a workspace. For details about how to create a Pangu user, see **Creating a Pangu User**.

1. Log in to ModelArts Studio.
2. Go to the workspace to which you want to add a user. In the navigation pane, choose **Workspace Management**. Click the **Member Management** tab.
3. Assign a role (for example, model development engineer) to the user, as shown in **Figure 2-20**. Search for the user name in the search box, select a role from the drop-down list, and click **Add** on the right to add the user to the workspace.

Figure 2-20 Adding a member as a model development engineer



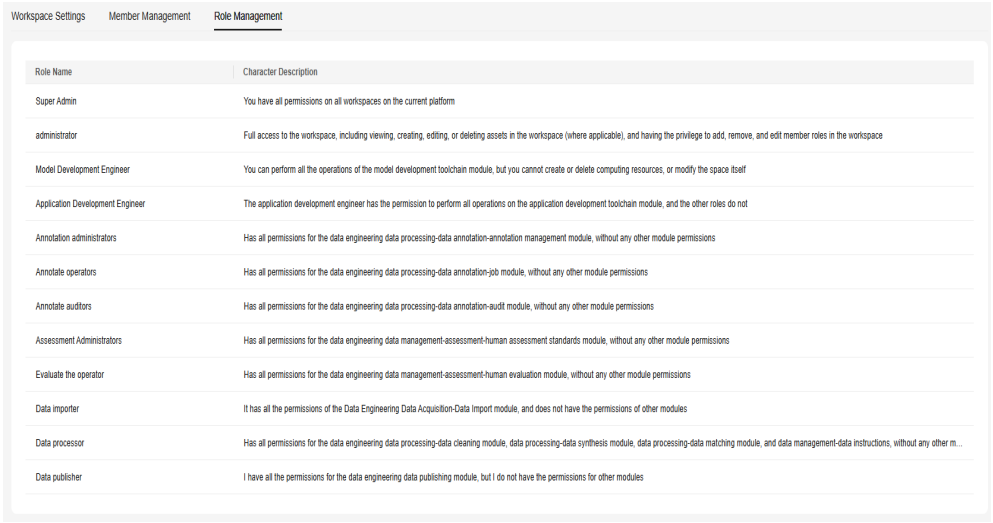
Modifying Pangu User Permissions

To modify the permissions of a user in a workspace, perform the following steps:

1. Log in to ModelArts Studio.
2. Go to the workspace whose user permissions need to be modified. In the navigation pane, choose **Workspace Management**. On the **Role Management** tab page, view the role name and permission description.

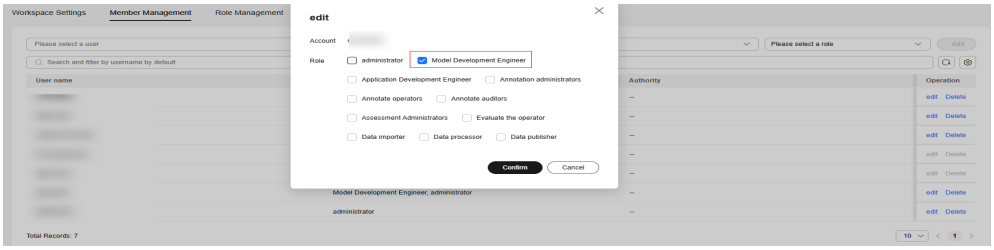
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 2-21 Role Management



3. Click the **Member Management** tab.
4. The following figure shows how to grant the model development engineer permission to a user. Click **edit** in the **Operation** column of the user list, select the role to be assigned to the user, and click **Confirm**.

Figure 2-22 Granting the Model Development Engineer Permission to a user



Removing a Pangu User

To delete a user from a workspace, perform the following steps:

1. Log in to ModelArts Studio.
2. Go to the workspace to which you want to delete a user. In the navigation pane, choose **Workspace Management**. Click the **Member Management** tab.
3. In the user list, locate the user that you want to delete and click **Delete** in the **Operation** column.

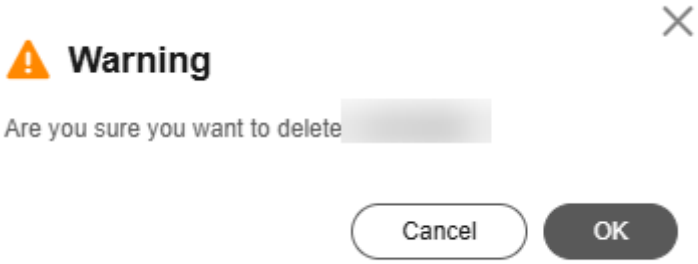
Figure 2-23 Member management



4. Click **OK** to confirm the deletion.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 2-24 Confirming the deletion



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3 Accessing Models in the Model Square

ModelArts Studio provides a one-stop model resource library on the **Model Square** page. You can select a proper model to execute training and inference jobs based on service requirements.

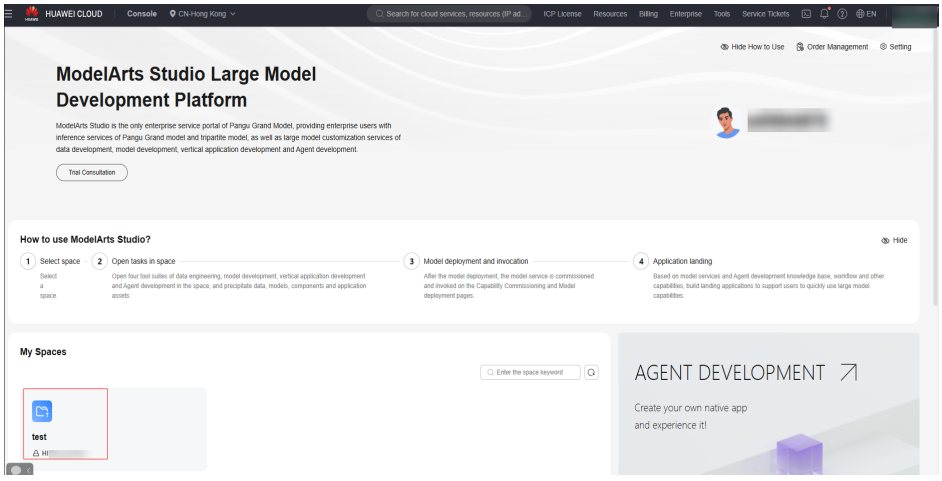
In the left pane of the **Model Square** page, you can select different filter criteria to filter the required models, or search for models by keyword. The right pane displays the models in cards, including the model name, model description, and application scenario tags. You can click **View Details** to go to the model details page and view the model version, description, and details. You can also click **Train** or **Deploy** to go to the model training or deployment page and create a training or deployment job for the model.

For details about the large models supported by the platform, see "Model Capabilities and Specifications" > "Third-Party Large Models" in *Service Overview*.

Accessing the Model Square

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 3-1 My Spaces

2. In the navigation pane, choose **Model Square**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3. In the left pane of the **Model Square** page, filter models by model type, model series, application scenario, supported function, and subscription type.

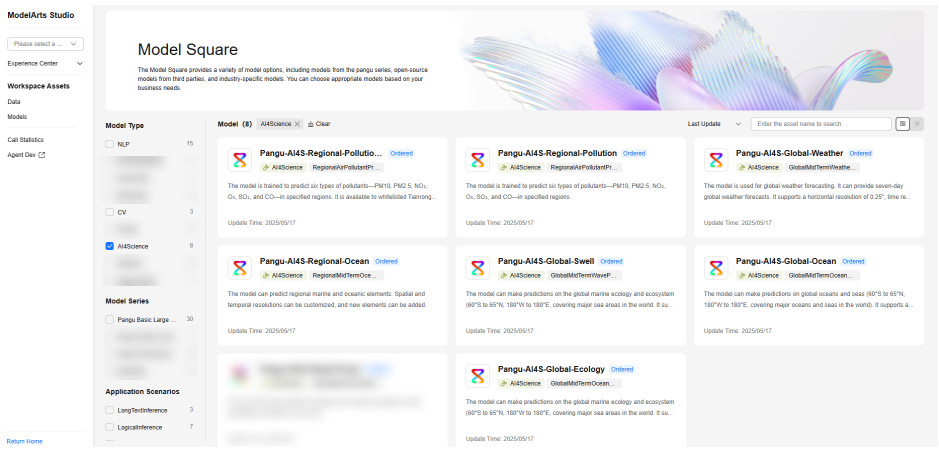
Table 3-1 Model filtering

Filter Criteria	Description
Model Type	Model types include **NLP**, **CV**, **AI4Science**, etc.
Model Series	You can select **Pangu Basic Large Model**.
Application Scenarios	The options include **LongTextInference**, **LogicalInference**, **SemanticSegmentation**, **ObjectDetection**, **DeepThinking**, **TextGeneration**, **DataAnalysis**, **IntentClassification**, etc.
Supported Functions	The options include **Support Training** and **Support Deployment**.
Subscription Type	The options include **Free** and **Paid**.

Viewing Model Version Information

On the **Model Square** page, move the pointer over the target model card and click **View Details**. On the model details page that is displayed, view the model details.

Figure 3-2 Model Square



The model details page displays the model introduction and version list. The **Model Introduction** tab page displays the model description, disclaimer, and platform service agreement. The **Version List** tab page displays the features of multiple versions of the current model.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 3-3 Model Introduction

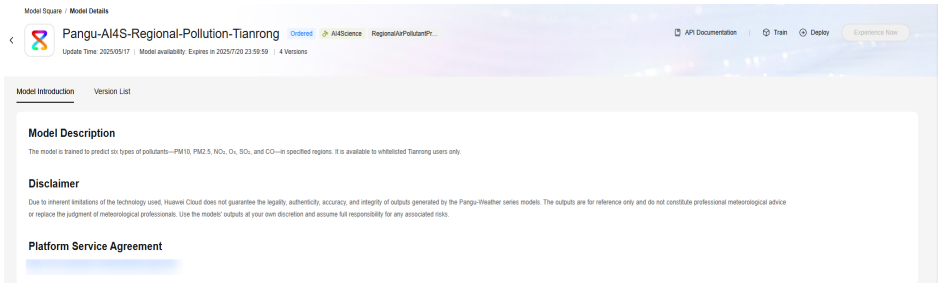
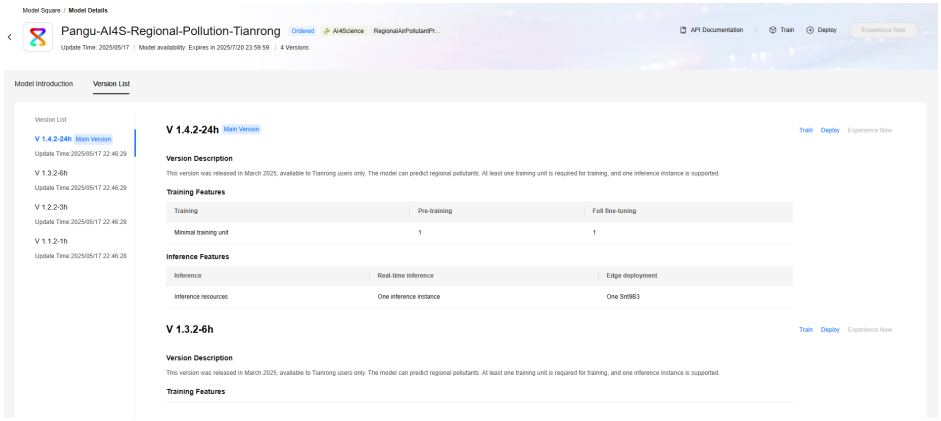


Figure 3-4 Version List



Training and Deploying a Model

For details about how to train and deploy a model, see the sections of the corresponding model.

Trying Out Model Functions

Deploy an existing model. The platform provides preset deployment services. You can click **Experience Center** to try out the model functions. For models that do not provide preset services, you need to deploy them on the **Model Development** > **Model Deployment** page.

1. In the navigation pane, choose **Model Development** > **Model Deployment**. Click **Create Deployment Task** in the upper right corner.
2. On the displayed page, select the model source and model type, set deployment parameters, and start model deployment.

   After the deployment is complete, the **Experience Now** button is displayed on the card of the corresponding model on the **Model Square** page. After you click the button, the **Experience Center** page is displayed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4 Using Data Engineering to Create a Dataset

4.1 Introduction to Data Engineering

What Is Data Engineering?

Data engineering is a one-stop data processing and management feature provided by ModelArts Studio Large Model Development Platform. Its goal is to facilitate the efficient and accurate utilization of data for large-scale model training through data acquisition, processing, and publishing. This feature aids in efficient data management and processing, enhances data quality and processing efficiency, and lays a robust data foundation for large model development.

Data engineering provides the following functions:

- **Data acquisition**: Data acquisition is the first step of data engineering. Data from different sources and in different formats can be imported to the platform. An original dataset can be generated.
  - Data can be imported through OBS.
  - The following data types are supported: **Text**, **Image**, **Video**, **Audio**, **Weather**, and **Other**.
  - Custom format: You can flexibly upload data in custom formats based on service requirements.

  With these functions, you can easily import a large amount of data to the platform to prepare for subsequent operations such as data processing and model training.
- **Data processing**: The platform provides data processing, data synthesis, data labeling, and data combination to ensure that raw data meets various service requirements and model training standards. A processed dataset can be generated.
  - Data processing: Data processing aims to preprocess data by using dataset processing operators. Dedicated processing operators are designed for different types of datasets to ensure that data meets model training standards and service requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

  - Data synthesis: Using either a preset or custom data instruction, process the original data, and generate new data based on a specified number of epochs.
  - Data labeling: Data labeling aims to add accurate labels to unlabeled datasets. The quality of labeled data directly affects the training effect and accuracy of models. The platform supports manual labeling and AI pre-labeling for different datasets.

    Image captions and video captions support AI pre-labeling.
  - Data combination is a process of combining multiple datasets into a processed dataset based on a specific ratio. A proper ratio ensures the diversity, balance, and representativeness of datasets.

  Through data processing, the platform can effectively clear noise data and standardize data formats, helping improve the overall quality of datasets.
- **Data Publishing:** The platform allows you to publish datasets in different modalities and formats and generates published datasets.
  - Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations.

    The following dataset formats are supported: standard format and Pangu format (applicable to Pangu model training). Currently, only text and image datasets can be published in Pangu format.
- **Data Management:** The platform provides dataset management and data evaluation to manage datasets of different types, ensure that data meets the diversity, balance, and representativeness requirements of large model training through data quality evaluation, and promote efficient data circulation and application.
  - **Datasets**: The platform manages datasets of different types, such as original datasets, processed datasets, and published datasets.
  - **Data Evaluation**: Data evaluation checks the quality of datasets and evaluates multiple dimensions of data based on evaluation standards to detect and solve potential problems.

In addition to data acquisition, data processing, and data publishing, the platform supports one-stop management of original datasets, processed datasets, published datasets, and data synthesis instructions. When building large-scale datasets, the data engineering capabilities of ModelArts Studio offer significant flexibility and efficiency, enabling seamless collaboration in data processing workflows and rapid adaptation to evolving service and technical requirements.

Data Types Supported by the Platform

Table 4-1 lists the data types supported by ModelArts Studio. For details about the data format requirements of each type, see **Dataset Format Requirements**.

Table 4-1 Data types supported by the platform

Data Type	Content	Supported File Format
Text	Document	txt, mobi, epub, docx, and pdf
	Web page	html

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Data Type	Content	Supported File Format
	Pre-trained text	jsonl
	Single-turn Q&A	jsonl and csv
	Single-turn Q&A (with a system persona)	jsonl and csv
	Multi-turn Q&A	jsonl
	Multi-turn Q&A (with a system persona)	jsonl
	Q&A ranking	jsonl and csv
	Direct Preference Optimization (DPO)	jsonl
	DPO (with a system persona)	jsonl
Image	Image Only	jpg, jpeg, png, bmp, and tar
	Image + Caption	• Supported image formats: JPG, JPEG, PNG, and BMP. All images must be saved as TAR packages. • Captions can be in JSONL format. JSONL files support only UTF-8 encoding.
	Image + QA Pair	• Supported image formats: JPG, JPEG, PNG, and BMP. All images must be saved as TAR packages. • QA pair format: JSONL. JSONL files support only UTF-8 encoding.
	Object detection	• Supported image formats: JPG, JPEG, PNG, BMP, TIF, and TIFF • Annotation file format: XML
	Image classification	• Supported image formats: JPG, JPEG, PNG, BMP, TIF, and TIFF • Annotation file format: TXT

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Data Type	Content	Supported File Format
	Anomaly detection	• Supported image formats: JPG, JPEG, PNG, and BMP • Annotation file format: TXT
	Semantic segmentation	• Image + XML – Supported image formats: JPG, JPEG, PNG, and BMP – Annotation file format: XML • Original image + annotated image + JSON – Supported image formats: JPG, JPEG, PNG, and BMP – Annotation file format: annotated image + JSON • Original image + Labeled image + TXT – Supported image formats: JPG, JPEG, PNG, and BMP – Annotation file format: annotated image + TXT • Original image + annotated image: The original and annotated images can be in JPG, JPEG, PNG, or BMP format. • Image + PNG
	Pose estimation	• Image + JSON – Supported image formats: JPG, JPEG, PNG, and BMP – Annotation file format: JSON • Image + XML – Supported image formats: JPG, JPEG, PNG, and BMP – Annotation file format: XML. Each image corresponds to an annotation file.
	Instance segmentation	• Supported image formats: JPG, JPEG, PNG, and BMP • Annotation file format: XML
	Change detection	• Supported image formats: JPG, JPEG, and BMP • Annotation file format: PNG
Video	Video	MP4 and AVI
	Video + Annotation	• Supported video formats: MP4 and AVI • Annotation file format: JSONL. JSONL files support only UTF-8 encoding.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Data Type	Content	Supported File Format
	Video classification	Supported file format: video + TXT. Video format: MP4 and AVI. Annotation file format: TXT. Each video corresponds to an annotation file.
	Event detection	• Supported video formats: MP4 and AVI. The duration of each video is greater than or equal to 128s. The FPS is greater than or equal to 10. • Annotation file format: JSON. A video can corresponds to one or multiple annotation files.
Audio	Audio Only	mp3, flac, wav, opus, aac, and m4a All audio files can be stored in multiple folders. Each folder can contain audio files in different formats.
	Audio + Annotation	• Supported audio formats: mp3, flac, wav, opus, aac, and m4a • Annotation file format: JSONL Note: JSONL files support only the UTF-8 encoding.
Weather	Weather Data	nc, cdf, netcdf, gr, gr1, grb, grib, grb1, grib1, gr2, grb2, and grib2
Other	Customization	You can customize dataset types based on specific scenarios.

Operations Supported by Each Type of Data

Table 4-2 lists the data engineering operations supported by each type of data.

Table 4-2 Operations supported by each type of data

Data Type	Data Acquisition	Data Processing	Data Synthesis	Data Labeling	Data Combination	Data Evaluation	Data Publishing
Text	√	√	√	√	√	√	√
Image	√	√	-	√	√	√	√
Audio	√	√	-	√	-	-	√
Weather	√	√	-	-	-	-	√
Other	√	-	-	-	-	-	√

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.2 Process of Using Data Engineering

High-quality data is the foundation for continuous iteration and optimization of large models. Its quality directly determines the performance, generalization capability, and adaptability of models to different application scenarios. Only through systematic preparation and processing of data can valuable insights be extracted, thereby providing better support for model training. Therefore, data acquisition, cleaning, labeling, evaluation, and publishing are indispensable steps in data development.

For details about the data engineering process, see **Figure 4-1** and **Table 4-3**.

Figure 4-1 Dataset construction flowchart

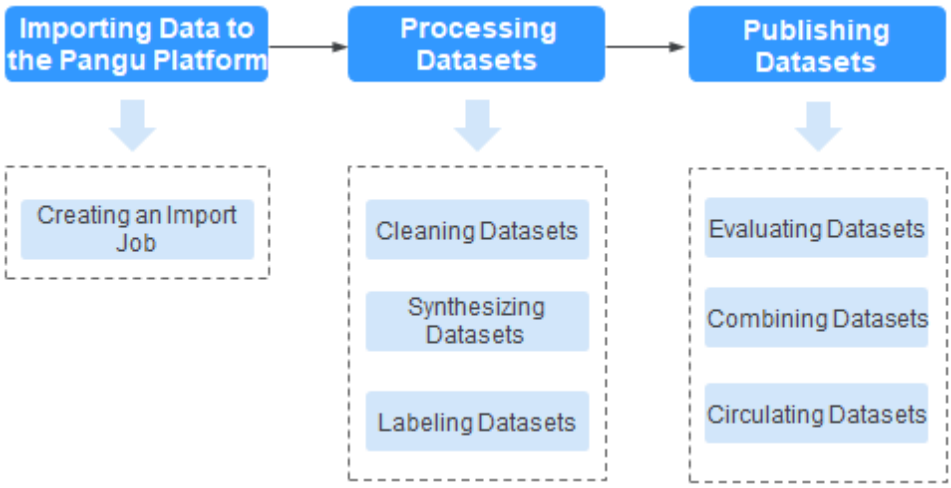


Table 4-3 Dataset construction process

Procedure	Step	Description
Importing data to the Pangu platform	Creating an import task	Import data stored in OBS or local data into the platform for centralized management, facilitating subsequent processing or publishing.
Processing datasets	Processing datasets	Use dedicated processing operators to preprocess data, ensuring it meets the model training standards and service requirements. Different types of datasets utilize operators specially designed for removing noise and redundant information, to enhance data quality.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Procedure	Step	Description
	Generating a synthetic dataset	Using either a preset or custom data instruction, process the original data, and generate new data based on a specified number of epochs. This process can extend the dataset to some extent and enhance the diversity and generalization capability of the trained model.
	Labeling datasets	Add accurate labels to unlabeled datasets to ensure high-quality data required for model training. The platform supports both manual annotation and AI pre-annotation. You can choose an appropriate annotation method based on your needs. The quality of data labeling directly impacts the training effectiveness and accuracy of the model.
	Combining datasets based on a specific ratio	Dataset combination involves combining multiple datasets based on a specific ratio and generating a processed dataset. A proper ratio ensures the diversity, balance, and representativeness of datasets and avoids issues resulting from uneven data distribution.
Publishing datasets	Evaluating datasets	The platform offers predefined evaluation standards for multiple types of data. You can choose from these predefined standards or customize evaluation standards as needed to precisely improve data quality, ensure that data meets high standards, and enhance model performance.
	Publishing datasets	Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations. Datasets can be published in standard or Pangu format. • Standard format: default dataset format on the platform. Datasets in this format can be published as assets, but cannot be used in the development of Pangu models. • Pangu format: When training a Pangu model, you need to publish it in this format. Currently, only text and image datasets can be published in Pangu format.

4.3 Dataset Format Requirements

4.3.1 Format Requirements for Text Datasets

ModelArts Studio supports the creation of text datasets. During the creation, you can import data in various formats. **Table 4-4** lists the format requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-4 Format requirements for text datasets

File Content	File Format	File Requirements
Document	txt, mobi, epub, docx, and pdf	Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.
Web page	html	Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.
Pre-trained text	jsonl	• JSONL format: **text** indicates the text data used for pre-training. The following is an example: `{"text":"Pangu Models are Pangu series AI models launched by Huawei, including the NLP model, multimodal model, CV model, scientific computing model, and prediction model."}` • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.

Single-turn Q&A	jsonl and csv	• JSONL format: The data consists of Q&A pairs. **context** and **target** indicate the question and answer, respectively. The following is an example:
		`{"context": "Hello, please introduce yourself.","target": "I am a Pangu model."}`
		• CSV format: The first column in the CSV file corresponds to **context**, and the second column corresponds to **target**. The following is an example:
		`"Hello, please introduce yourself.","I am a Pangu model."`
		• Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited.
		Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Single-turn Q&A (with a system persona)	jsonl and csv	• JSONL format: **system** indicates the persona, **context** indicates the question, and **target** indicates the answer. {"system": "You're a smart and humorous Q&A assistant.","context": "Hello, please introduce yourself.","target":"Hello. I'm your smart assistant."} • CSV format: In the CSV file, the first column corresponds to **system**, and the second and third columns correspond to **context** and **target**, respectively. "You're a smart and humorous Q&A assistant.","Hello, please introduce yourself.","Hello. I'm your smart assistant." • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.
Multi-turn Q&A	jsonl	• JSONL format: an array consisting of at least one Q&A pair. The format is [{"context":"context content 1","target":"target content 1"}, {"context":"context content 2","target":"target content 2"}]. **context** and **target** indicate the question and answer, respectively. [{"context":"Hello","target":"Hello, what can I do for you?"}, {"context":"Please introduce Huawei Cloud products.","target":"Huawei Cloud products include but are not limited to compute, storage, and network products."}] • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.
Multi-turn Q&A (with a system persona)	jsonl	• JSONL format: an array consisting of at least one Q&A pair. **system** indicates the persona, **context** indicates the question, and **target** indicates the answer. [{"system": "You are a book recommendation expert."}, {"context":"Hi","target":"Hi. What can I do for you?"}, {"context":"Can you recommend some books to me?","target":"Of course. Based on your interest, I recommend you the Future of Autonomous Driving."}] • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Q&A ranking	jsonl and csv	• JSONL format: **context** indicates the question. The order of **targets** 1, 2, and 3 represent the order of human-preferred answers. The most preferred answer is placed at the forefront. {"context":"context content ","targets":["Answer 1,""Answer 2,""Answer 3"]} • CSV format: The first column in the CSV file corresponds to **context**, and the other columns are answers. "Question,""Answer 1","Answer 2,","Answer 3" • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.
Direct Preference Optimization (DPO)	jsonl	• JSONL format: **context** indicates the question, **target** indicates the expected correct answer, and **bad_target** indicates an incorrect or unexpected answer. Single-turn Q&A {"context": ["Hello, please introduce yourself."],"target":"I'm a Pangu model.","bad_target":"Sorry, I can't assist with that."} Multi-turn Q&A {"context": ["Hello, please introduce yourself.", "I'm a Pangu model.", "Please introduce Huawei Cloud products."],"target":"Huawei Cloud products include but are not limited to compute, storage, and network products."} • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

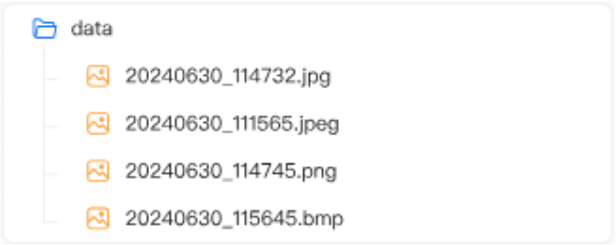
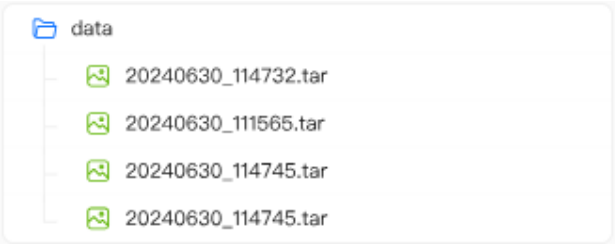
File Content	File Format	File Requirements
DPO (with a system persona)	jsonl	• JSONL format: **system** indicates the persona, **context** indicates the question, **target** indicates the expected correct answer, and **bad_target** indicates an incorrect or unexpected answer. Single-turn Q&A (with a system persona) {"system":"You are a humorous Q&A assistant.","context":["Hello, please introduce yourself."],"target":"Hello. I am your smart assistant. How can I help you?","bad_target":"Sorry, I can't assist with that."} Multi-turn Q&A (with a system persona) {"system":"You are a humorous Q&A assistant.","context":["Hello, please introduce yourself.", "Hello, I am your smart assistant. How can I help you?", "Please introduce Huawei Cloud products."], "target":"Huawei Cloud provides compute, storage, and network products, as well as many other products.","bad_target":"Sorry, I can't assist with that." • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.

4.3.2 Format Requirements for Image Datasets

ModelArts Studio supports the creation of image datasets. During the creation, you can import data in various formats. **Table 4-5** lists the format requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

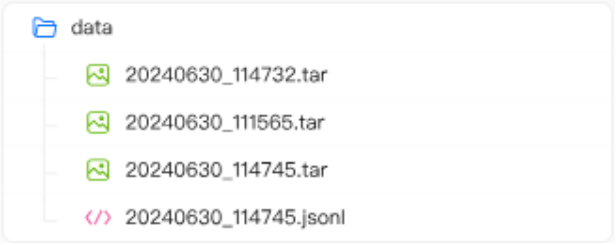
Table 4-5 Image dataset format requirements

File Content	File Format	File Requirements
Image only	TAR and image directory	• Image: JPG, JPEG, PNG, and BMP  • TAR: The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.  • Import from OBS: The size of a single compressed package cannot exceed 50 GB (only .tar packages are supported). The size of a single file cannot exceed 50 GB. The number of files is not limited. Local upload: The size of a single compressed package cannot exceed 10 MB (only .tar packages are supported). The size of a single file cannot exceed 10 MB. A maximum of 100 files are supported.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Image + Caption	Image: TAR; Caption: JSONL	• Image: TAR. Multiple TAR packages are supported. The TAR package stores original images. Each image name must be unique, for example, abc.jpg. Image: JPG, JPEG, PNG, and BMP • JSONL: The image description JSONL file is stored in the outermost directory. One TAR package corresponds to one JSONL file. Each line in the file content represents a segment of text. The format is as follows: `{"image_name":"Image name (abc.jpg)","tar_name":"TAR package name (1.tar)","caption":"Text description of the image"}` • Import from OBS: The size of a single compressed package cannot exceed 50 GB (only .tar packages are supported). The size of a single file cannot exceed 50 GB. The number of files is not limited. Local upload: The size of a single compressed package cannot exceed 10 MB (only .tar packages are supported). The size of a single file cannot exceed 10 MB. A maximum of 100 files are supported. 

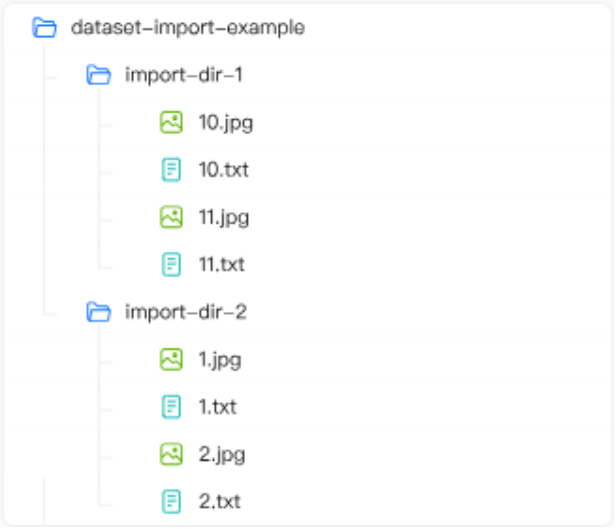
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Image + QA Pair	Image: TAR; QA pair: JSONL	• Image: TAR. Multiple TAR packages are supported. The TAR package stores original images. Each image name must be unique, for example, abc.jpg. Image: JPG, JPEG, PNG, and BMP • JSONL: The image description JSONL file is stored in the outermost directory. One TAR package corresponds to one JSONL file. Each line in the file content represents a segment of text. The format is as follows: `{"image_name":"Image name (abc.jpg)","tar_name":"TAR package name (1.tar)","conversations":[{"question":"Question 1","answer":"Answer 1"},{"question":"Question 2","answer","Answer 2"}]}` • Import from OBS: The size of a single compressed package cannot exceed 50 GB (only .tar packages are supported). The size of a single file cannot exceed 50 GB. The number of files is not limited. Local upload: The size of a single compressed package cannot exceed 10 MB (only .tar packages are supported). The size of a single file cannot exceed 10 MB. A maximum of 100 files are supported. 

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Object detection	PASCAL VOC	• The dataset consists of image files and corresponding annotation files. The annotation files must be in PASCAL VOC format. Labeled objects and their annotation files (mapped to the labeled objects) must be in the same directory. For example, if the name of the labeled object file is **IMG_2.jpg**, the name of the annotation file must be **IMG_2.xml**. • Images can be in JPG, JPEG, PNG, BMP, TIF, or TIFF format. Annotation files must be in XML format. For details, see **Specifications of Annotation Files in an Object Detection Dataset**. • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. 

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Image classification	Image + TXT	• The dataset consists of image files and corresponding annotation files. Labeled objects and their annotation files (mapped to the labeled objects) must be in the same directory. • Images can be in JPG, JPEG, PNG, BMP, TIF, or TIFF format. Annotation files must be in TXT format. For details, see **Description of an Annotation File for an Image Classification Dataset**. • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. 

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	File Requirements
Instance segmentation	Image + XML	• The file storage mode must meet the format required by Segment Anything/Instance Segmentation. • Supported image formats: JPG, JPEG, PNG, and BMP; Supported annotation file format: XML. • Annotations use bounding boxes in the PASCAL VOC format. Annotations and images must have the same name and must be stored in the same folder. • For details about annotation files in XML format, see **Description of an Annotation File for an Instance Segmentation Dataset**. • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. 

Specifications of Annotation Files in an Object Detection Dataset

The following description follows the annotation file format for object detection in **Table 4-5**.

The object detection dataset supports annotation files in ModelArts PASCAL VOC 1.0 format.

Labeled objects and their annotation files (in one-to-one relationship with the labeled objects) must be in the same directory. For example, if the name of the labeled object file is **IMG_20180919_114745.jpg**, the name of the annotation file must be **IMG_20180919_114745.xml**.

The annotation files must be in PASCAL VOC format, a standardized XML annotation format used for labeling image datasets. A PASCAL_VOC file contains information on the image directory, image file code, image size, and object information. For details about the format, see **Table 4-6**.

Example of a file uploaded to OBS:

```
dataset-import-example
    IMG_20180919_114732.jpg
    IMG_20180919_114732.xml
    IMG_20180919_114745.jpg
    IMG_20180919_114745.xml
    IMG_20180919_114945.jpg
    IMG_20180919_114945.xml
```

An XML annotation file example is as follows:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<annotation>
   <folder>NA</folder>
   <filename>bike_1_1593531469339.png</filename>
   <source>
      <database>Unknown</database>
   </source>
   <size>
      <width>554</width>
      <height>606</height>
      <depth>3</depth>
   </size>
   <segmented>0</segmented>
   <object>
      <name>Dog</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
      <occluded>0</occluded>
      <bndbox>
         <xmin>279</xmin>
         <ymin>52</ymin>
         <xmax>474</xmax>
         <ymax>278</ymax>
      </bndbox>
   </object>
   <object>
      <name>Cat</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
      <occluded>0</occluded>
      <bndbox>
         <xmin>279</xmin>
         <ymin>198</ymin>
         <xmax>456</xmax>
         <ymax>421</ymax>
      </bndbox>
   </object>
</annotation>
```

Table 4-6 PASCAL VOC format description

Field	Mandatory (Yes/No)	Description
folder	Yes	Name of the directory where the image is located
filename	Yes	Name of the labeled file
size	Yes	Image pixel • **width**: image width. This parameter is mandatory. • **height**: image height. This parameter is mandatory. • **depth**: number of image channels. This parameter is mandatory.
segmented	Yes	Segmented or not. The value can be **0** or **1**. The value **0** means no segmentation, and **1** means segmentation.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Field	Mandatory (Yes/No)	Description
object	Yes	Target object information, which includes the category, pose, truncation status, identification difficulty, and bounding box of an object. An image may contain more than one object. • **name**: type of the labeled object. This parameter is mandatory. • **pose**: shooting angle of the labeled object. This parameter is mandatory. • **truncated**: This parameter is mandatory. The value can be **0** or **1**. The value **0** indicates that the labeled object is truncated, and **1** indicates the opposite. • **occluded**: This parameter is mandatory. The value can be **0** or **1**. The value **0** indicates that the labeled object content is occluded, and **1** indicates the opposite. • **difficult**: This parameter is mandatory. The value can be **0** or **1**. The value **0** indicates that the labeled object is easy to recognize, and **1** indicates the opposite. • **confidence**: This parameter is optional. The value ranges from **0** to **1**. A value closer to 1 indicates a higher level of confidence. • **bndbox**: bounding box type. This parameter is mandatory. For details about the possible values, see **Table 4-7**.

Table 4-7 Bounding box types

type	Shape	Labeling Information
point	Point	Coordinates of a point <x>100<x> <y>100<y>

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

type	Shape	Labeling Information
line	Line	Coordinates of points <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>
bndbox	Rectangle	Coordinates of the upper left and lower right points <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>
polygon	Polygon	Coordinates of points <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	Circle	Center coordinates and radius <cx>100<cx> <cy>100<cy> <r>50<r>

Description of an Annotation File for an Image Classification Dataset

The following description follows the annotation file format for image classification in **Table 4-5**.

The image classification dataset supports annotation files in ModelArts image classification 1.0 format.

Labeled objects and their annotation files (in one-to-one relationship with the labeled objects) must be in the same directory. An annotation file in TXT format can contain a single label or multiple labels.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- The image and annotation files must be stored in the same directory, with the content in the annotation file used as the label of the image.

  In the following example, **import-dir-1** and **import-dir-2** are the imported subdirectories.

  ```
  dataset-import-example
  ├─import-dir-1
  │      10.jpg
  │      10.txt
  │      11.jpg
  │      11.txt
  │      12.jpg
  │      12.txt
  └─import-dir-2
         1.jpg
         1.txt
         2.jpg
         2.txt
  ```

  The following shows an annotation file for a single label, for example, the **1.txt** file:

  ```
  Cat
  ```

  The following shows an annotation file for multiple labels, for example, the **2.txt** file:

  ```
  Cat
  Dog
  ```

Specifications of Annotation Files in an Anomaly Detection Dataset

The following description follows the annotation file format for anomaly detection in **Table 4-5**.

The labeling files and images must be stored in the same folder.

- The image and annotation files must be stored in the same directory, with the content in the annotation file used as the label of the image (normal or abnormal).

  Example:

  ```
  dataset-import-example
  │      IMG_20180919_114732.jpg
  │      IMG_20180919_114732.txt
  │      IMG_20180919_114745.jpg
  │      IMG_20180919_114745.txt
  ```

  The following shows an annotation file for the "abnormal" label, for example, the **IMG_20180919_114732.txt** file:

  ```
  abnormal
  ```

  The following shows an annotation file for the "normal" label, for example, the **IMG_20180919_114745.txt** file:

  ```
  normal
  ```

Description of JSON Annotation Files for a Posture Estimation Dataset

The following description follows the annotation file format for post estimation in **Table 4-5**.

Posture estimation dataset labeling is based on the open-source character keypoint labeling format (COCO). The **annotations**, **train**, and **val** folders must be included. In the **annotations** folder, **train.json** and **val.json** contain the

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

annotations of the training set and validation set. The **train** and **val** folders store images. The following is an example:

```
├─annotations
│      train.json
│      val.json
├─train
│      IMG_20180919_114745.jpg
├─val
│      IMG_20180919_114945.jpg
```

The following is an example of a JSON annotation file:

```
{
   "images": [
      {
         "license": 2,
         "file_name": "000000000139.jpg",
         "coco_url": "",
         "height": 426,
         "width": 640,
         "date_captured": "2013-11-21 01:34:01",
         "flickr_url": "",
         "id": 139
      }
   ],
   "annotations": [
      {
         "num_keypoints": 15,
         "area": 2913.1104,
         "iscrowd": 0,
         "keypoints": [
            427,
            170,
            1,
            429,
            169,
            2,
            0,
            0,
            0,
            434,
            168,
            2,
            0,
            0,
            0,
            441,
            177,
            2,
            446,
            177,
            2,
            437,
            200,
            2,
            430,
            206,
            2,
            430,
            220,
            2,
            420,
            215,
            2,
            445,
            226,
            2,
            452,
            223,
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
            2,
            447,
            260,
            2,
            454,
            257,
            2,
            455,
            290,
            2,
            459,
            286,
            2
         ],
         "image_id": 139,
         "bbox": [
            412.8,
            157.61,
            53.05,
            138.01
         ],
         "category_id": 1,
         "id": 230831
      },
   ],
   "categories": [
      {
         "supercategory": "person",
         "id": 1,
         "name": "person",
         "keypoints": [
            "nose",
            "left_eye",
            "right_eye",
            "left_ear",
            "right_ear",
            "left_shoulder",
            "right_shoulder",
            "left_elbow",
            "right_elbow",
            "left_wrist",
            "right_wrist",
            "left_hip",
            "right_hip",
            "left_knee",
            "right_knee",
            "left_ankle",
            "right_ankle"
         ],
         "skeleton": [
            [
               16,
               14
            ],
            [
               14,
               12
            ],
            [
               17,
               15
            ],
            [
               15,
               13
            ],
            [
               12,
               13
```

```
            ],
            [
              6,
              12
            ],
            [
              7,
              13
            ],
            [
              6,
              7
            ],
            [
              6,
              8
            ],
            [
              7,
              9
            ],
            [
              8,
              10
            ],
            [
              9,
              11
            ],
            [
              2,
              3
            ],
            [
              1,
              2
            ],
            [
              1,
              3
            ],
            [
              2,
              4
            ],
            [
              3,
              5
            ],
            [
              4,
              6
            ],
            [
              5,
              7
            ]
          ]
        }
    ]
}
```

Table 4-8 COCO format description

Field	Mandatory (Yes/No)	Description
images	Yes	Image information.
license	No	License identifier of an image.
file_name	Yes	Image file name.
coco_url	No	URL of an image in the official COCO dataset.
height	Yes	Image height (in pixels).
width	Yes	Image width (in pixels).
date_captured	No	Date and time when an image is captured.
flickr_url	No	URL of an image on the Flickr website.
id	Yes	Unique identifier of an image.
annotations	Yes	Labeling information.
num_keypoints	Yes	Number of labeled key points.
area	Yes	Area of the bounding box, in pixel squares.
iscrowd	Yes	Whether the scenario is a complex group scenario (for example, crowded people). The value **0** indicates that the scenario is not a crowded scenario, and the value **1** indicates that the scenario is a crowded scenario.
keypoints	Yes	Coordinates and visibility of labeled key points. All key points are listed in sequence. Each key point is represented by three numbers: [x, y, v]. **x** and **y** are pixel coordinates of the key point, and **v** is visibility (**0**: invisible and not in the image; **1**: invisible but in the image; **2**: visible and in the image).
image_id	Yes	ID of the image associated with the annotation. The value must be the same as the value of **id** in the **images** field.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Field	Mandatory (Yes/No)	Description
bbox	Yes	Bounding box of the target object, represented by [x, y, width, height], where **x** and **y** are the coordinates of the upper left corner of the bounding box, and **width** and **height** are the width and height of the bounding box.
category_id	Yes	ID of a label category. For human posture estimation, the value is usually **1** (indicating person).
id	Yes	Unique identifier of an image.
categories	Yes	Label type information.
supercategory	Yes	Upper-level category of a category, which is usually **person**.
id	Yes	Unique identifier of a category, usually **1** for human posture estimation.
name	Yes	Name of a category, which is usually **person**.
keypoints	Yes	List of key point names. Generally, 17 key points are defined in the COCO format, such as nose, left_eye, right_eye, left_ear, right_ear, left_shoulder, right_shoulder, left_elbow, right_elbow, left_wrist, right_wrist, left_hip, right_hip, left_knee, right_knee, left_ankle, and right_ankle.
skeleton	Yes	List of skeleton connections, which are used to indicate the connection relationships between key points. Each connection is represented by a pair of key point indexes, for example, [1, 2], indicating a connection line from a nose (nose) to a left eye (left_eye).

Description of an Annotation File for an Instance Segmentation Dataset

The following description follows the annotation file format for instance segmentation in **Table 4-5**.

Labeled objects and their annotation files (in one-to-one relationship with the labeled objects) must be in the same directory. For example, if the name of the labeled object file is **IMG_20180919_114745.jpg**, the name of the annotation file must be **IMG_20180919_114745.xml**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

The annotation files must be in PASCAL VOC format, a standardized XML annotation format used for labeling image datasets. A PASCAL_VOC file contains information on the image directory, image file code, image size, and object information. For details about the format, see **Table 4-9**.

Example of a file uploaded to OBS:

```
├─dataset-import-example
│      IMG_20180919_114732.jpg
│      IMG_20180919_114732.xml
│      IMG_20180919_114745.jpg
│      IMG_20180919_114745.xml
```

Example of an XML annotation file:

```
<annotation>
 <folder>NA</folder>
 <filename>0001.jpg</filename>
 <source>
  <database>Unknown</database>
 </source>
 <size>
  <width>2560</width>
  <height>1440</height>
  <depth>3</depth>
 </size>
 <segmented>1</segmented>
 <mask_source></mask_source>
 <object>
  <name>aggregate</name>
  <pose>Unspecified</pose>
  <truncated>0</truncated>
  <difficult>0</difficult>
  <mask_color>238,130,238</mask_color>
  <occluded>0</occluded>
  <polygon>
   <x1>657.0</x1>
   <y1>357.0</y1>
   <x2>645.0</x2>
   <y2>351.0</y2>
   <x3>624.0</x3>
   <y3>352.0</y3>
   <x4>616.0</x4>
   <y4>353.0</y4>
  </polygon>
 </object>
</annotation>
```

Table 4-9 PASCAL VOC format description

Field	Mandatory (Yes/No)	Description
folder	Yes	Name of the directory where the image is located
filename	Yes	Name of the labeled file

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Field	Mandatory (Yes/No)	Description
size	Yes	Image pixel • **width**: image width. This parameter is mandatory. • **height**: image height. This parameter is mandatory. • **depth**: number of image channels. This parameter is mandatory.
segmented	Yes	Segmented or not. The value can be **0** or **1**. The value **0** means no segmentation, and **1** means segmentation.
object	Yes	Target object information, which includes the category, pose, truncation status, identification difficulty, and bounding box of an object. An image may contain more than one object. • **name**: type of the labeled object. This parameter is mandatory. • **pose**: shooting angle of the labeled object. This parameter is mandatory. • **truncated**: This parameter is mandatory. The value can be **0** or **1**. The value **0** indicates that the labeled object is truncated, and **1** indicates the opposite. • **occluded**: This parameter is mandatory. The value can be **0** or **1**. The value **0** indicates that the labeled object content is occluded, and **1** indicates the opposite. • **difficult**: This parameter is mandatory. The value can be **0** or **1**. The value **0** indicates that the labeled object is easy to recognize, and **1** indicates the opposite. • **confidence**: This parameter is optional. The value ranges from **0** to **1**. A value closer to 1 indicates a higher level of confidence. • **bndbox**: bounding box type. This parameter is mandatory. For details about the possible values, see **Table 4-10**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-10 Bounding box types

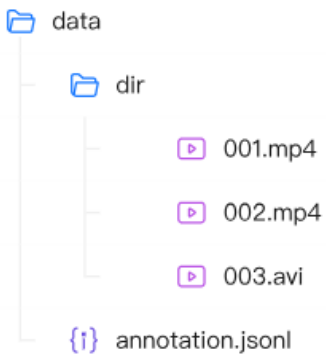
type	Shape	Labeling Information
point	Point	Coordinates of a point <x>100<x> <y>100<y>
line	Line	Coordinates of points <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>200<y2>
bndbox	Rectangle	Coordinates of the upper left and lower right points <xmin>100<xmin> <ymin>100<ymin> <xmax>200<xmax> <ymax>200<ymax>
polygon	Polygon	Coordinates of points <x1>100<x1> <y1>100<y1> <x2>200<x2> <y2>100<y2> <x3>250<x3> <y3>150<y3> <x4>200<x4> <y4>200<y4> <x5>100<x5> <y5>200<y5> <x6>50<x6> <y6>150<y6>
circle	Circle	Center coordinates and radius <cx>100<cx> <cy>100<cy> <r>50<r>

4.3.3 Format Requirements for Video Datasets

ModelArts Studio supports the creation of video datasets. During the creation, you can import data in various formats. **Table 4-11** lists the format requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-11 Format requirements for video datasets

File Content	File Format	Requirement
Video	MP4 or AVI	• Videos in MP4 and AVI formats can be uploaded. Videos can be stored in multiple folders, and each folder can contain videos in MP4 or AVI format. • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited.
Video + Annotation	Video + JSON	• The video format can be MP4 and AVI. • Annotation files must be in JSONL format. The encoding format can only be UTF-8. The following is an example.  For details about the annotation file in JSONL format, refer to the following: {"video_fn": "13/ad098173-af09-48fe-95c3-e72fd629688e.mp4" Relative path of the video, "prompt": "A person pours a clear liquid from a bottle into a shot glass, then lifts the glass to their mouth and drinks the shot. The background includes a red coat and other indistinct background elements." Video synopsis generation (simplified), "long_prompt": "A person is seen pouring a clear liquid from a green glass bottle into a small glass. The individual is wearing a white shirt with a lace collar and a beige cardigan. The background appears to be a cozy indoor setting, possibly a cafe or a restaurant, with red and white elements visible, such as a red coat hanging on the wall and a white table. The person carefully pours the liquid, ensuring it is filled to the brim of the glass. The liquid is clear and has some green leaves floating in it. The person then holds the glass up, possibly to show the contents or to prepare for a drink.", Video synopsis generation (detailed) }

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

<table>
<tr><th>File Content</th><th>File Format</th><th>Requirement</th></tr>
<tr><td>Event detection</td><td>Video + JSON</td><td>Data source samples must be in AVI or MP4 format, and annotation files must be in JSON format. The data source must contain two or more files in AVI or MP4 format. One video file can correspond to one or more annotation files.
The duration of each video must be greater than 128s, the FPS must be greater than or equal to 10, and the test set and training set must contain videos.
Supported video formats include MP4 and AVI. The duration of each video must be greater than 128s, and the FPS must be greater than or equal to 10. The annotation.json file is the annotation file.
Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited.
The following is an example.
Video + JSON (many-to-one)
 
Video + JSON (one-on-one)
 
For details about the annotation file in JSON format, refer to the following:
<pre>{
'version': 'dataset_name_v.x.x',// Dataset version information.
'classes': [category1',category2', ...],// List of all category names. Each category corresponds to a label, which is used to mark events or actions in the video.
'database': {
'video_name':{
// Training set: train; test set: test.
'subset': 'train',
'duration': 1660.3, // Total video duration, in seconds.
'fps': 30.0,// Video frame rate.
'width': 720,// Video width, in pixels.
'height': 1280,// Video height, in pixels.
'ext': 'mp4',//Video file name extension.
// 34.5 and 42.4 indicate the start time and end time, respectively. The unit is</pre></td></tr>
</table>

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	Requirement

		second.
		// label indicates the category. It must be an element in the classes list, indicating the event or action type corresponding to the video clip.
		'annotations': [
		{'label': 'category1', 'segment': [34.5, 42.4]},
		{'label': 'category1', 'segment': [124.4, 142.9]},
		...
		]
		},
		'video_name':{
		'subset': xxx,// Video file name, excluding the file name extension.
		'duration': xxx,
		'fps': xxx,
		'width': xxx,
		'height': xxx,
		'ext': xxx,
		'annotations': [
		{'label': xxx, 'segment': xxx},
		{'label': xxx, 'segment': xxx},
		...
		]
		},
		...
		}
		}

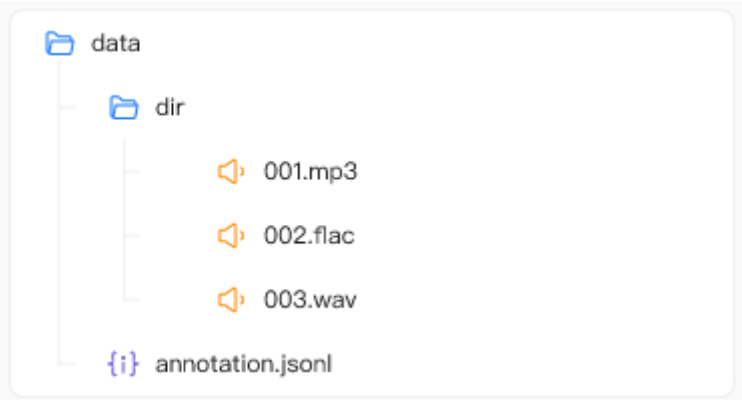
Video classification	Video + TXT	• Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. • The video format can be MP4 or AVI, and the annotation file format must be TXT. Each video corresponds to an annotation file. The following is an example. 

4.3.4 Format Requirements for Audio Datasets

ModelArts Studio supports the creation of audio datasets. During the creation, you can import data in various formats. **Table 4-12** lists the format requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-12 Format requirements for audio datasets

File Content	File Format	Requirement
Audio Only	Audio	• Supported formats: mp3, flac, wav, opus, aac, and m4a. All audio files can be stored in multiple folders. Each folder can contain audio files in different formats. • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited.
Audio + Annotation	Audio + JSONL	• Supported audio formats: mp3, flac, wav, opus, aac, and m4a Annotation file format: JSONL • Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. The following is an example.  For details about the annotation file in JSONL format, refer to the following: {"audio_name":"dir/16k_16bit_1channel_2s.flac","caption":"1"} {"audio_name":"dir/16k_16bit_1channel_2s.mp3","caption":"2"} {"audio_name":"dir/16k_16bit_1channel_2s.opus","caption":"3"} {"audio_name":"dir/16k_16bit_1channel_2s.wav","caption":"4"}

4.3.5 Format Requirements for Weather Datasets

ModelArts Studio supports the import of weather datasets, including **marine meteorological** data.

Marine meteorological data is usually derived from weather reanalysis. Weather reanalysis uses modern weather models and data assimilation technologies to reprocess historical observation data and generate high-quality weather records. The data can be global or region-specific, aiming to provide complete, consistent, and high-resolution weather data.

The reanalysis data is in binary format. For details about the format requirements, see **Table 4-13**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-13 Format requirements for weather datasets

File Content	File Format	Example
Meteorology - Weather data	nc, cdf, netcdf, gr, gr1, grb, grib, grb1, grib1, gr2, grb2, and grib2	• Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Weather data usually includes global or regional weather variables, such as temperature (T), atmospheric pressure (P), and wind speed (U and V). In a file, these variables may be arranged by time, geographical range, and pressure level. The following is an example: {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["P", "T", "U", "V"], "upper_air_layers": ["1000hPa", "100hPa", "150hPa", "175hPa", "200hPa", "250hPa", "300hPa", "400hPa", "500hPa", "50hPa", "600hPa", "700hPa", "850hPa", "925hPa"], "upper_air_features": ["Q", "T", "U", "V", "Z"]} • **geo_range**: defines the geographical range covered by data. The value of **lat** ranges from -90.0 to 90.0, and the value of **lon** ranges from 0.0 to 360.0. • **time_range**: time range of data. The timestamp is in milliseconds. • **total_size**: total size of the files to download, in bytes. • **surface_features**: list of surface variables, such as atmospheric pressure (P), temperature (T), and wind speed (U and V). • **upper_air_layers**: upper-air pressure levels, for example, 1000 hPa and 100 hPa. • **upper_air_features**: upper-air variables, for example, humidity (Q), temperature (T), wind speed (U and V), and height (Z).

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	Example
Meteorology - Ocean data	nc, cdf, netcdf, gr, gr1, grb, grib, grb1, grib1, gr2, grb2, and grib2	• Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. • Ocean data usually contains global or regional ocean variables, such as temperature (T), atmospheric pressure (P), and wind speed (U and V). The following is an example of the format: {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["SSH", "T", "P", "U", "V"], "under_sea_layers": ["0m", "6m", "10m", "20m", "30m", "50m", "70m", "100m", "125m", "150m", "200m", "250m", "300m", "400m", "500m"], "under_sea_features": ["T", "U", "V", "S"]} – **geo_range**: defines the geographical range covered by data. The value of **lat** ranges from -90.0 to 90.0, and the value of **lon** ranges from 0.0 to 360.0. – **time_range**: time range of data. The timestamp is in milliseconds. – **total_size**: total size of the files to download, in bytes. – **surface_features**: list of sea surface variables, such as sea surface height (SSH), temperature (T), and wind speed (U and V). – **under_sea_layers**: deep-sea pressure levels, for example, 500 m and 400 mPa. – **under_sea_features**: list of deep-sea variables, such as sea salt (S), temperature (T), and sea flow rate (U and V).
Meteorology - Ecological data	nc, cdf, netcdf, gr, gr1, grb, grib, grb1, grib1, gr2, grb2, and grib2	• Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Ecological data usually includes ecological variables such as total chlorophyll concentration (Tca), chlorophyll concentration (Chl), and diatom concentration (Dia). The following is an example: {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["Tca","Chl","Dia","Coc","Cya","Irn","Nit","MLD"]} • **geo_range**: defines the geographical range covered by data. The value of **lat** ranges from -90.0 to 90.0, and the value of **lon** ranges from 0.0 to 360.0. • **time_range**: time range of data. The timestamp is in milliseconds. • **total_size**: total size of the files to download, in bytes. • **surface_features**: ecological variables, for example, Tca, Chl, and Dia.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

File Content	File Format	Example
Meteorology - Wave data	nc, cdf, netcdf, gr, gr1, grb, grib, grb1, grib1, gr2, grb2, and grib2	• Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited. Wave data usually includes the significant wave height (SWH). The following is an example: {"geo_range": {"lat": ["-90.0", "90.0"], "lon": ["0.0", "360.0"]}, "time_range": ["1640995200000", "1641164400000"], "total_size": 7376211808, "surface_features": ["SWH"]} • **geo_range**: defines the geographical range covered by data. The value of **lat** ranges from -90.0 to 90.0, and the value of **lon** ranges from 0.0 to 360.0. • **time_range**: time range of data. The timestamp is in milliseconds. • **total_size**: total size of the files to download, in bytes. • **surface_features**: wave variable, which is significant wave height (SWH).

4.3.6 Format Requirements for Other Datasets

In addition to text, image, video, weather, and prediction datasets, the platform supports the import of other types of datasets, that is, user-defined datasets used for model training.

Other types of datasets support only the operations in **Processing Other Datasets** and **Publishing Other Datasets**.

Import from OBS: The size of a single file cannot exceed 50 GB, and the number of files is not limited.

Local upload: The size of a single file cannot exceed 10 MB, and the number of files cannot exceed 100.

4.4 Importing Data to the Pangu Platform

A dataset is a set of related data samples used for processing and analysis.

After data stored in OBS or local data is imported to ModelArts Studio, an original dataset is generated and managed by ModelArts Studio for subsequent processing or publishing.

Creating an Import Job

Before creating an import job, prepare data based on **Dataset Format Requirements**.

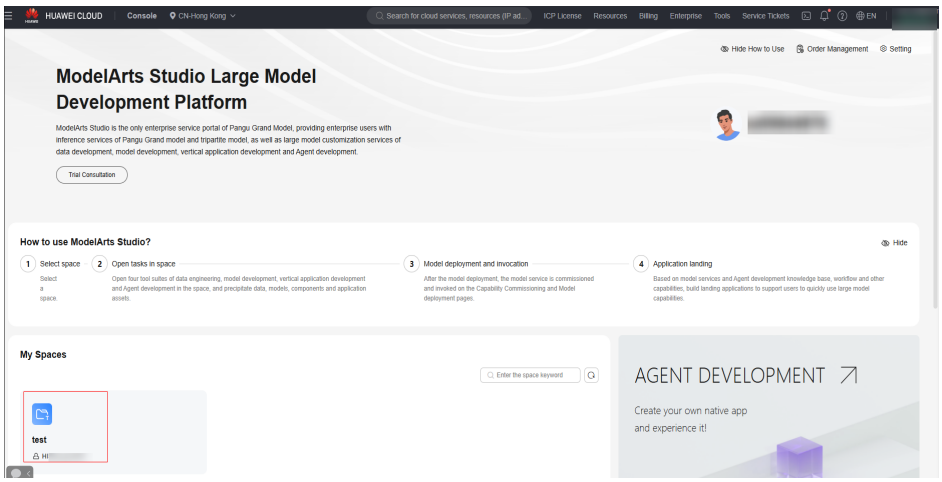
You can use OBS to import data. For details, see **Using OBS Console**.

To create an import job, do as follows:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

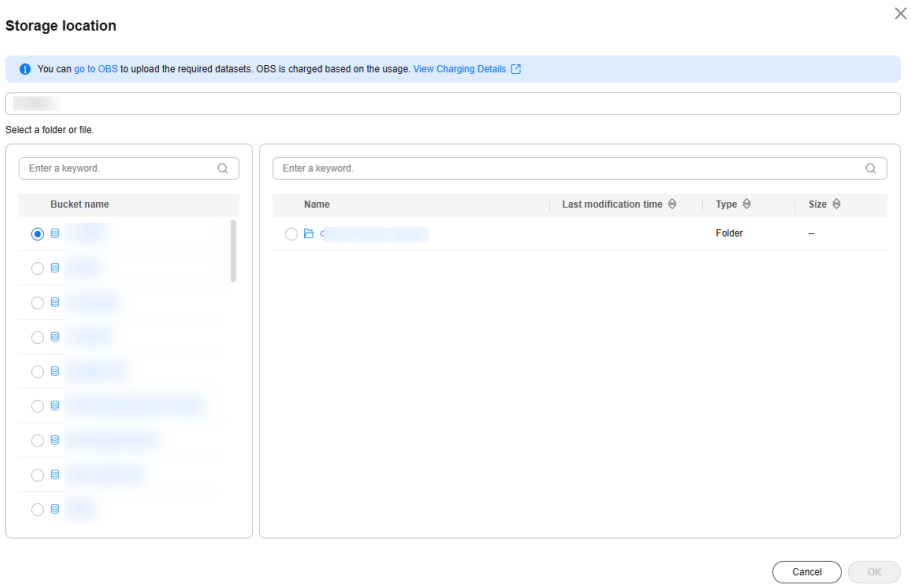
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 4-2 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Acquisition**. On the **Import Task** page, click **Create Import Job** in the upper right corner.
3. On the **Create Import Job** page, select the dataset type, file format, and import source.
4. Set **Import Source** to **OBS** and click  . In the **Storage Location** dialog box, select the data to be imported and click **OK**.

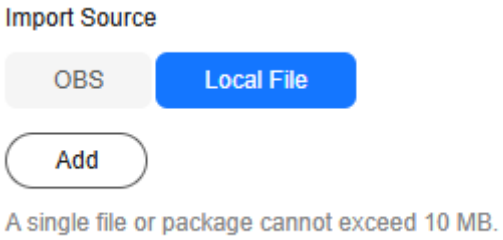
Figure 4-3 Selecting data to be imported



Set **Import Source** to **Local File** and click **Add**. Select the file to be imported.

Figure 2 Local File

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



5. Enter the dataset name and description. Enter extended information if required.

   Extended Info includes **Dataset Property** and **Dataset Copyright**.

   - **Dataset Property**: You can add industry, language, and custom information to a dataset.
   - **Dataset Copyright**: In addition to users' self-built datasets, open-source datasets may be used for model training. The dataset copyright function is used to record and manage the copyright information of datasets to ensure that data is used in compliance with laws and regulations and clearly understand the dataset sources and related copyright authorization. By filling in the information, you can trace the source of the data and specify the restrictions and permissions for using the data, thereby protecting data copyright and avoiding copyright disputes.

6. Click **Create Now** in the lower right corner of the page to return to the **Import Task** page. On the page that is displayed, you can view the task status of the dataset. If the task status is **Succeeded**, the data is successfully imported.
7. To view the imported dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Original Dataset** tab.

   NOTE

   If the task status is **Failed**, the import has failed. The possible causes are as follows:

   - The file name extension is incorrect. Check whether the file name extension is correct. For example, if you create a dataset in CSV format, the file name extension must be .csv.
   - The file content fails to be verified. Check whether the format of the uploaded file is correct. You can download data samples on the **Create Import Job** page for comparison.

Managing Original Datasets

After data is imported, you can manage the original datasets in a unified manner. You can view the basic information, data lineage, and operation records of the dataset, and download and delete the dataset.

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Datasets**. On the **Original Dataset** tab page, click the name of the dataset to be viewed.
   - Enter the basic information about the dataset. On the **Basic Information** tab page, you can view the data details, data source, and extended information.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- In the **Extended Info** area, you can set dataset properties as required, including the dataset property name, industry, language, and custom tag.
- Download the original dataset. On the **Data Preview** tab page, you can view the data content and click **Download** in the upper right corner to download the original dataset.
- View data lineages. On the **Data Lineage** tab page, you can view the complete operations performed on the current dataset, such as processing and labeling.
- View operation records. On the **Operation Record** tab page, you can view the operation records of the current dataset, such as the creation time, status, and operator of the dataset.

3. Delete the original dataset. Click **Delete** in the **Operation** column. In the displayed dialog box, click **OK**.

   NOTE

   Deleting an original dataset is a high-risk operation. Before deleting a dataset, ensure that it is no longer used.

4.5 Processing Datasets

4.5.1 Dataset Processing Scenarios

Introduction to Data Processing

ModelArts Studio provides the data processing function, covering data processing, data synthesis, and data labeling. This function ensures that the original data meets service requirements and model training standards. This function is the core of data engineering.

- **Data processing**

  Use dedicated processing operators to preprocess data, ensuring it meets the model training standards and service requirements. Different types of datasets utilize operators specially designed for removing noise and redundant information, to enhance data quality. In addition, you can create custom operators to flexibly process data based on specific service scenarios and model requirements. This further optimizes the data processing process and improves the accuracy and robustness of models.

- **Data synthesis**

  Using either a preset or custom data instruction, process the original data, and generate new data based on a specified number of epochs. This process can extend the dataset to some extent and enhance the diversity and generalization capability of the trained model.

- **Data labeling**

  Add accurate labels to unlabeled datasets to ensure high-quality data required for model training. The platform supports both manual annotation and AI pre-annotation. You can choose an appropriate annotation method based on your needs. The quality of data labeling directly impacts the training effectiveness and accuracy of the model.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- **Data combination**

  Dataset combination involves combining multiple datasets based on a specific ratio into a processed dataset. A proper ratio ensures the diversity, balance, and representativeness of datasets and avoids issues resulting from uneven data distribution.

Through data processing, the platform can effectively clear noise data and standardize data formats, helping improve the overall quality of datasets. Data processing aims at optimization based on data types and service scenarios to provide high-quality input for model training and improve model performance.

Significance of Data Processing

Data processing plays an important role in large model development. It provides the following benefits:

- **Improved data quality**

  Raw data usually contains noise, missing values, or inconsistency, which directly affects the model training effect. Through data processing, invalid information can be effectively removed and missing data can be filled to ensure data accuracy and consistency, thereby improving data quality and providing reliable input for model training.

- **Improved diversity and generalization capabilities of datasets**

  When the data volume is insufficient or samples are unbalanced, data synthesis can generate new data and expand the scale and diversity of datasets. By increasing data diversity, the generalization capabilities of the model in various scenarios can be improved, and the adaptability of the model to unknown data can be enhanced.

- **Enhanced effectiveness of model training**

  High-quality data is the basis for model training. Data processing aims at targeted optimization based on data types and business requirements to make data more compliant with training standards and improve training efficiency and model accuracy.

- **Ensured alignment with business requirements**

  Different business scenarios and model applications have different requirements on data. Data processing involves customized processing of data based on specific business requirements to ensure that data meets application scenario requirements, improve the alignment between data and the model, and improve the accuracy of business decision-making and the model accuracy.

- **Higher efficiency**

  With the automated data processing function provided by the platform, you can efficiently preprocess a large amount of data. This reduces manual intervention, improves data processing consistency and efficiency, and ensures smooth operation of the entire data engineering process.

- **Ensured data quality and adaptation**

  Data combination based on a specified ratio ensures that datasets meet the high requirements of large model training. This not only includes the requirements on the data scale, but also covers the data quality, balance, and representativeness. This way, data balance or diversity can be ensured, and the accuracy and robustness of the model can be improved.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- **Improved data diversity and representativeness**

  You can combine multiple datasets based on a specific ratio to ensure the diversity and representativeness of datasets in different task scenarios. This avoids excessive bias towards a certain type of data, ensures that the model can learn multiple features, and improves the adaptability to various situations.

In general, this function not only improves data handling efficiency, but also supports efficient model training by optimizing data quality and targeted processing. Through data processing, you can quickly build high-quality datasets and promote the successful development of large models.

Dataset Types That Support Data Processing

Table 4-14 lists the dataset types that support data processing.

Table 4-14 Dataset types that support data processing

Dataset Modality	Dataset Type	Data Processing	Data Synthesis	Data Labeling	Data Combination
Text	Documents	√	-	-	√
	Web pages	√	-	-	√
	Pre-trained text	√	√	-	√
	Single-turn Q&A	√	√	√	√
	Single-turn Q&A (with a system persona)	√	√	√	√
	Multi-turn Q&A	√	-	√	√
	Multi-turn Q&A (with a system persona)	√	-	√	√
	Q&A ranking	√	-	√	√
	Direct Preference Optimization (DPO)	-	-	-	√
	DPO (with a system persona)	-	-	-	√
Image	Image	√	-	√	√
	Image + Caption	√	-	√	√

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Dataset Modality	Dataset Type	Data Processing	Data Synthesis	Data Labeling	Data Combination
	Image + QA Pair	√	-	√	√
	Object detection	-	-	-	-
	Image classification	-	-	-	-
	Anomaly detection	-	-	-	-
	Semantic segmentation	-	-	-	-
	Pose estimation	-	-	-	-
	Instance segmentation	-	-	-	-
	Change detection	-	-	-	-
Video	Videos	√	-	√	-
	Video + Annotation	-	-	-	-
	Video content comprehension	√	-	√	-
	Event detection	-	-	-	-
	Video classification	-	-	-	-
Audio	Audio Only	√	-	√	-
	Audio + Annotation	-	-	-	-
Weather	Weather Data	√	-	-	-
Prediction	Time series (classification)	-	-	-	√
	Time series (regression)	-	-	-	-
	Structured (classification)	-	-	-	√
	Structured (regression)	-	-	-	-

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Dataset Modality	Dataset Type	Data Processing	Data Synthesis	Data Labeling	Data Combination
Other	Other	√ (Only custom operators can be used for data processing.)	-	-	-

4.5.2 Processing Text Datasets

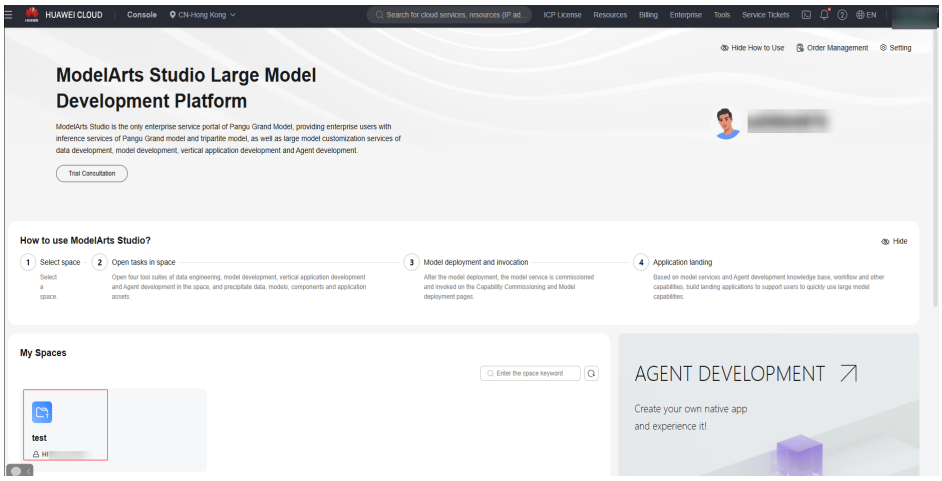
4.5.2.1 Processing Text Datasets

Before processing a text dataset, import data. For details, see **Importing Data to the Pangu Platform**.

To create a text dataset processing task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-4 My Spaces

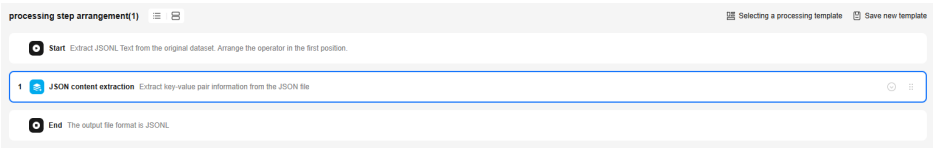
2. In the navigation pane, choose **Data engineering** > **Data Processing** > **Processing Tasks**. Click **Create Processing Job** in the upper right corner.
3. On the **Create Processing Job** page, select the text dataset to be processed and click **Next**.
4. Go to the **processing step arrangement** page. For details about available processing operators for text datasets, see **Text Dataset Processing Operators**.
   a. In the **Adding Operator** pane on the left, select the required operators. You can select one or more operators based on the actual scenario.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

The platform supports preset operators and custom operators. For details about how to create a custom operator, see **Custom Data Processing Operators**.

b. On the **processing step orchestration** page on the right, set operator parameters. You can drag ⠿ on the right to adjust the operator execution sequence.

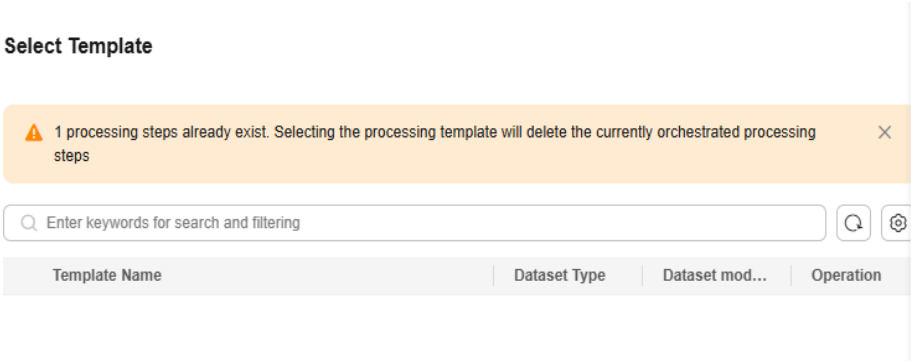
Figure 4-5 Operator orchestration



c. During orchestration, you can click **Save new template** in the upper right corner to save the current orchestration process as a template. During the creation of subsequent data processing tasks, you can select a processing template.

If you select a processing template, the orchestrated processing steps will be deleted.

Figure 4-6 Selecting a processing template



5. After the processing steps are orchestrated, click **Next** to go to the **Task Configuration** page.
   - Resource Allocation

     Click ⌄ to expand resource configuration and set task resources. You can also customize parameters. Click **Add Parameters** and enter the parameter name and value.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

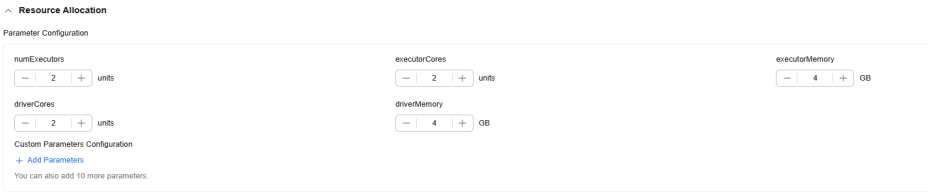
Table 4-15 Parameter configuration

Parameter Name	Description
numExecutors	Number of executors. The default value is 2. An executor is a process running on a worker node. It executes tasks and returns the calculation result to the driver. One core in an executor can run one task at the same time. Therefore, more tasks can be processed at the same time if you increase the number of the executors. You can add executors (if they are available) to process more tasks concurrently and improve efficiency. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16.
executorCores	Number of CPU kernels used by each executor process. The default value is 2. Multiple cores in an executor can run multiple tasks at the same time, which increases the task concurrency. However, because all cores share the memory of an executor, you need to balance the memory and the number of cores. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
executorMemory	Memory size used by each Executor process. The default value is 4. Specifies the executor memory used for job execution and communication. You can increase the memory for a job that requires a great number of resources, and run small jobs concurrently with a smaller memory. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
driverCores	Number of CPU kernels used by each driver process. The default value is 2. The driver schedules jobs and communicates with executors. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Description
driverMemory	Memory used by the driver process. The default value is 4. The driver schedules jobs and communicates with executors. Add driver memory when the number and parallelism level of the tasks increases. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.

Figure 4-7 Resource Allocation

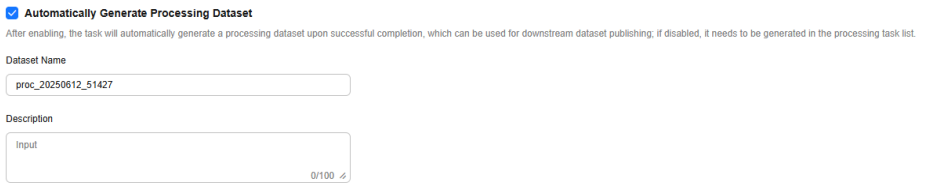


- Automatically Generate Processing Dataset

  Select and configure the information about the generated dataset, as shown in **Figure 4-8**. Click **OK** in the lower right corner. The platform starts the data combination task. After the task is successfully executed, a processed dataset is automatically generated.

  If you do not select this option, click **OK** in the lower right corner. The platform starts the combination task. After the combination task is successfully executed, manually generate a processed dataset.

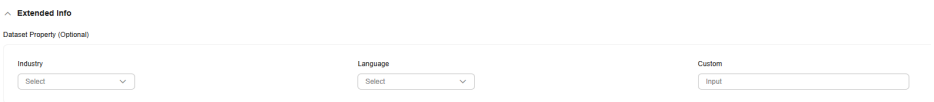
Figure 4-8 Automatically Generate Processing Dataset



- (Optional) Extended Info

  You can select the industry and language, or customize dataset properties.

Figure 4-9 Extended Info



6. Click **Start Process**. After the data processing job is successfully executed, the status changes from **Running** to **Succeeded**, indicating that the data has been processed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

NOTE

- After data processing is complete, if you do not need to use the data labeling and data synthesis functions, click **Generate** in the **Operation** column on the **Processing Tasks** page to generate a processed dataset.
- To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

4.5.2.2 Synthesizing Text Datasets

Data synthesis: Using either a preset or custom data instruction, process the original data, and generate new data based on a specified number of epochs.

Currently, the data synthesis function supports the synthesis of single-turn Q&A, single-turn Q&A (persona) Q&A ranking, DPO, and DPO (persona) data.

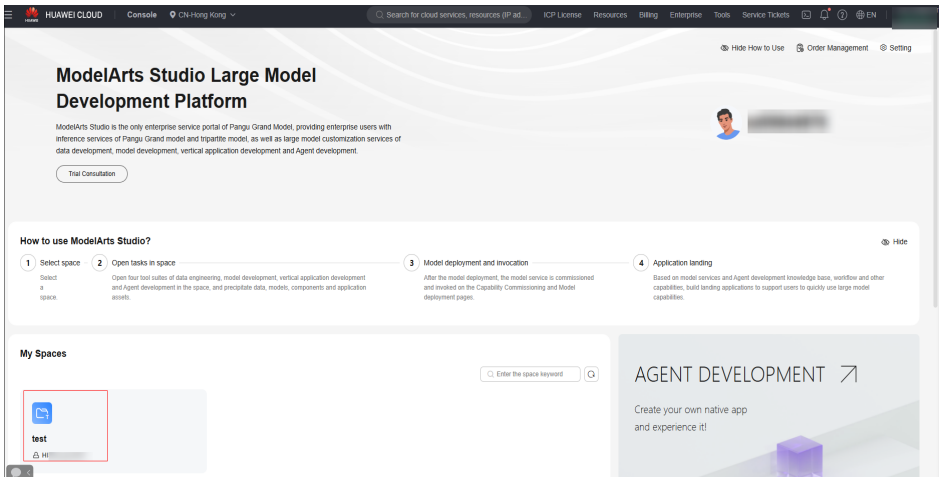
Creating a Text Data Synthesis Task

Before synthesizing a text dataset, import data. For details, see **Importing Data to the Pangu Platform**.

To create a text dataset synthesis task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

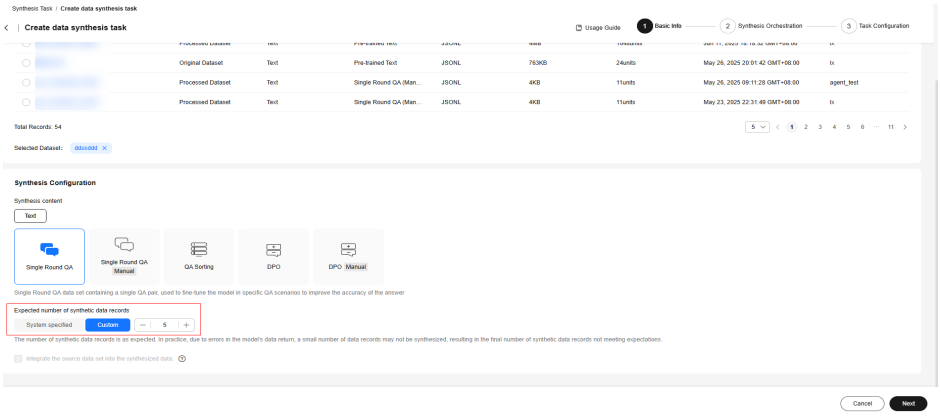
   Figure 4-10 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Synthesis Task**. On the displayed page, click **Create data synthesis task** in the upper right corner.
3. On the **Create data synthesis task** page, select the dataset to be synthesized, and select the content to be synthesized and the expected number of synthetic data records, as shown in **Figure 4-11**. The expected number of synthetic data records is the initial number of data records in the dataset.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-11 Expected number of synthetic data records



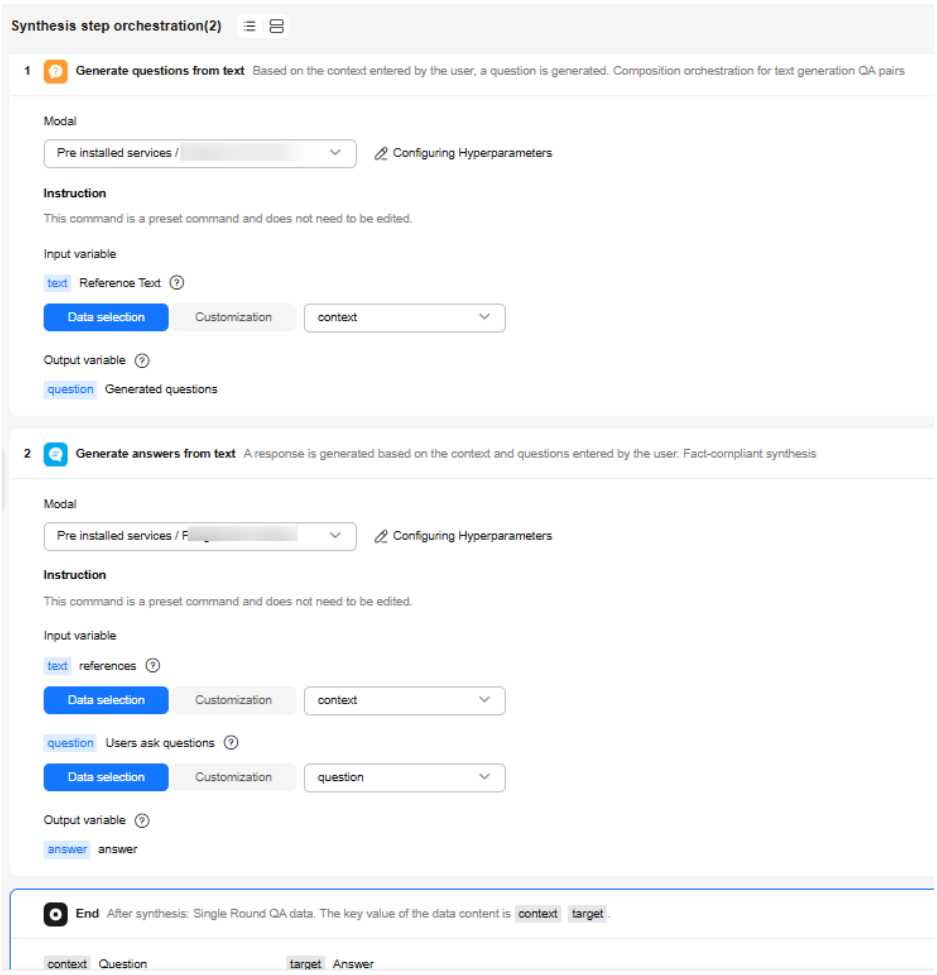
If the source dataset type is the same as the target dataset type, you can enable **Integrate the source data set into the synthesized data**. After the synthesis task is complete, the generated data is merged with the original dataset. Click **Next**.

4. On the **Synthesis Orchestration** page, orchestrate the synthesis steps.
   - In the **Add Instruction** pane, you can select a preset or custom instruction, or copy an instruction.
     - Preset instructions: The platform provides multiple preset instructions for you to execute synthesis tasks. For details, see **Introduction to Preset Data Instructions**.
     - Custom instructions: The platform supports orchestration of custom instructions. For details about how to create a custom instruction, see **Creating a Custom Data Synthesis Instruction**.
   - After selecting a command, click **OK** and set command parameters.

     Figure 4-12 shows an example of configuring synthesis instruction parameters for a pre-trained text dataset. The synthesis task uses the pre-trained text to generate Q&A pairs.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-12 Example of configuring parameters for combining pre-trained text data into an instruction



- After the instruction orchestration is complete, click **Enable commissioning** in the upper right corner to preview the instruction effect.

5. After the instruction commissioning is complete, click **Next** in the lower right corner. The **Task Configuration** page is displayed.
   - Select **Automatically Generate Processing Dataset** if required.

     Select this option and configure the information for generating a processed dataset, as shown in **Figure 4-13**. Click **Create and Start** in the lower right corner. The platform starts the data synthesis task. After the synthesis task is successfully executed, a processed dataset is automatically generated.

     If you do not select this option, click **Create and Start** in the lower right corner. The platform starts the data synthesis task. After the synthesis task is successfully executed, manually generate a processed dataset.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-13 Automatically Generate Processing Dataset



- Enter the dataset name, description, and extended information (optional, including the industry, language, and custom information).

6. After the data synthesis task is successfully executed, the status changes from **Running** to **Success**, indicating that the data has been synthesized.

NOTE

After data synthesis is complete, if you do not need to use the data labeling and data combination functions, click **Generate** in the **Operation** column on the **Synthesis Task** page to generate a processed dataset.

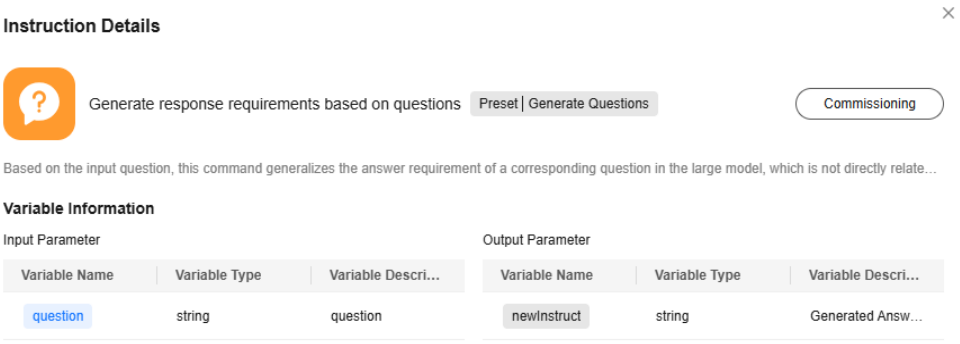
To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

Introduction to Preset Data Instructions

The data synthesis function of ModelArts Studio provides you with preset instructions. You can choose **Data Engineering** > **Data Processing** > **Data Synthesis**. On the **Synthesis Task** page, click **Manage Synthesis Instructions**. On the **System Preset** tab page, view instruction details, as shown in **Figure 4-14**. Click **Commissioning** to view the commissioning guide, as shown in **Figure 4-15**.

For details about the list of preset instructions, see **Table 4-16**.

Figure 4-14 Instruction details



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-15 Commissioning guide

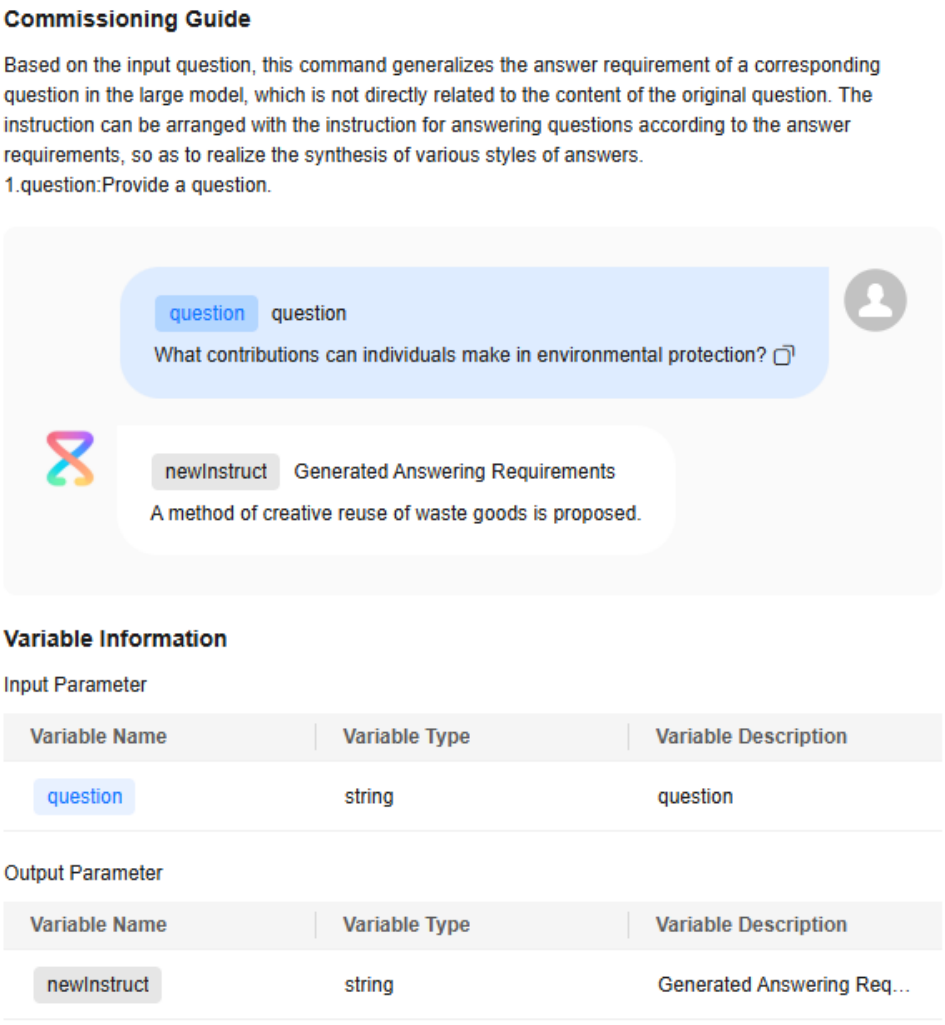


Table 4-16 List of preset instructions

Category	Instruction Name	Description
Generate Questions	Rewrite the question to a less difficult one.	This instruction can be used to enable the model to generate a simpler and less difficult question based on the question entered by the user.
	Rewriting the question to a more difficult one	This instruction can be used to enable the model to generate a more complex and more difficult question based on the question entered by the user.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Instruction Name	Description
	Generating answering requirements based on the questions	The instruction generalizes answer requirements of a corresponding question based on an input question, and the requirements are not directly related to content of the original question. The instruction may be orchestrated with an instruction for answering a question according to answering requirements, to implement synthesis of various styles of answers.
	Generating similar questions based on samples_few-shot	The instruction generates, by using a plurality of question examples entered by the user, one or more new questions that match the sample style.
	Generating questions based on text	This instruction generates a question based on the context entered by the user. It can be used to synthetize and orchestrate text generation Q&A pairs.
	Question rewriting	This instruction is used to rewrite a question to generate more complex questions, which can be used for instruction generalization.
Generate Answers	Answer rewriting	This instruction is used to modify the answer style based on the user-specified settings without changing the answer content. It can be used to generate and orchestrate instructions with persona settings to implement generalization of Q&A pairs.
	Generating answers based on text_complying with requirements	This instruction is used to generate answers based on the input context and the instructions and questions specified by users. It can be used to generate and orchestrate instructions to implement generalization of factual Q&A pairs.
	Generating an answer to a question	This instruction is used to generate answers based on questions.
	Generating an answer based on the text_with a designated persona	This instruction is used to generate answers based on the input context and user-specified persona settings and questions. It can be used to generate and orchestrate instructions with persona settings to implement generalization of factual Q&A pairs.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Instruction Name	Description
	Generating questions and answers based on questions_with a designated persona	This instruction is used to generate questions and answers based on the user-specified persona settings and questions. It can be used to generate and orchestrate instructions with persona settings to implement generalization of Q&A pairs.
	Generating answers based on text	This instruction is used to generate answers based on the context and questions entered by users. It can be used for synthesis of factual answers.
Generating Q&A pairs	Generating Q&A pairs based on text_true or false	This instruction is used to construct a true or false question from the text provided by the user, and provide a correct answer.
	Generating Q&A pairs based on text_fill-in-blank	This instruction is used to construct a fill-in-blank question from the text provided by the user, and provide a correct answer.
	Generating Q&A pairs based on text_single-answer question	This instruction is used to construct a single-answer question with four options from the text provided by the user, and provide a correct answer.
	Generating Q&A pairs based on text_multiple-answer question	This instruction is used to construct a multiple-answer question with four options from the text provided by the user, and provide a correct answer.
	Generating Q&A pairs based on text_Q&A question	This instruction is used to construct a question from the text provided by the user, and provide a correct answer.
	Extracting Q&A pairs based on text_financial scenario	This instruction is used to extract Q&A pairs based on financial documents entered by users.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Instruction Name	Description
Generating a persona setting	Generating a persona setting based on a question	This instruction is used to generate a persona setting based on the question entered by the user.
Other	Generalizing bad cases	This instruction is used to generate attack questions that may let the model make mistakes in similar scenarios based on the bad case questions and answers provided by users. Users can specify the number of attack questions to be generated. A maximum of 10 attack questions can be generated.
	Deriving the problem solving idea based on the answer	This instruction is used to generate the corresponding problem solving ideas based on the questions and answers entered by the user.
	Instruction generalization	This instruction is used to generalize instructions based on the style specified by the user. It can be used to generate and orchestrate related instructions that match Q&A pairs of the specified requirements to implement Q&A pair generalization.

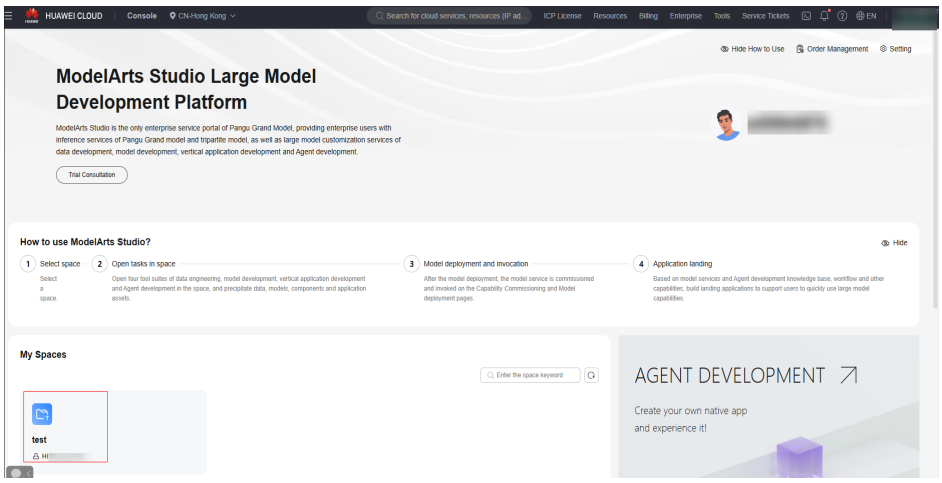
Creating a Custom Data Synthesis Instruction

The platform allows you to create custom data synthesis instructions.

This section uses the scenario of generating a topic prose passage as an example to describe how to configure a custom data synthesis instruction.

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 4-16 My Spaces

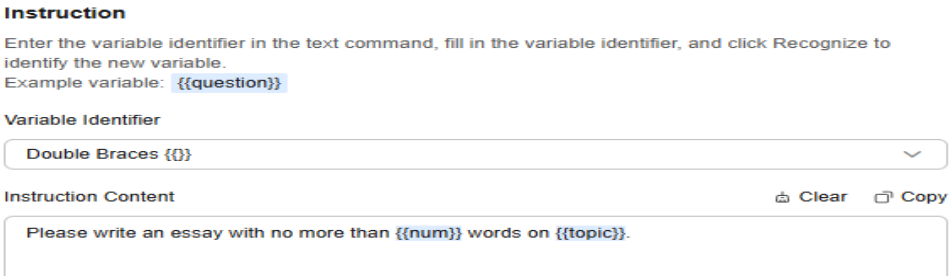


 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Synthesis**. On the displayed page, click **Manage Synthetic Instructions**. On the **Custom** tab page, click **Create Instruction**.
3. In the **Create Instruction** dialog box, enter the name and description, and click **OK**. The page for configuring a synthetic instruction is displayed.
4. Select the variable identifier "{{}}" and enter the instruction "Use {{topic}} as the subject and write a prose passage with no more than {{num}} characters."

   Click **Identify** and then click **Yes**.

Figure 4-17 Instruction configuration



5. Set variables based on **Table 4-17**.

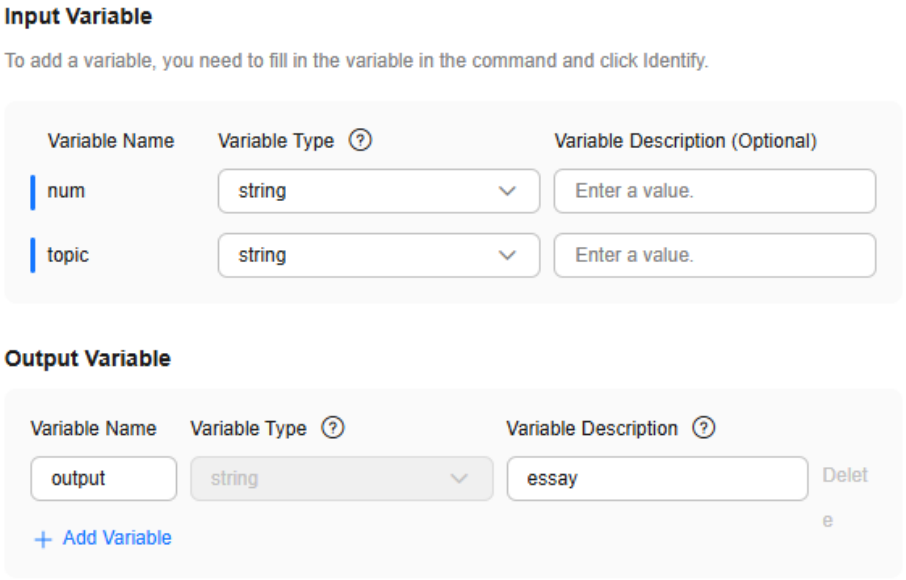
Table 4-17 Data instruction variable configuration

Variable Type	Variable Name	Data Type	Variable Description
Input variables	topic	string	Topic
	num	string	Word count
Output variables	output	string	Prose

The variable description field of an output variable is the content that can be understood by the large model and needs to be filled in carefully.

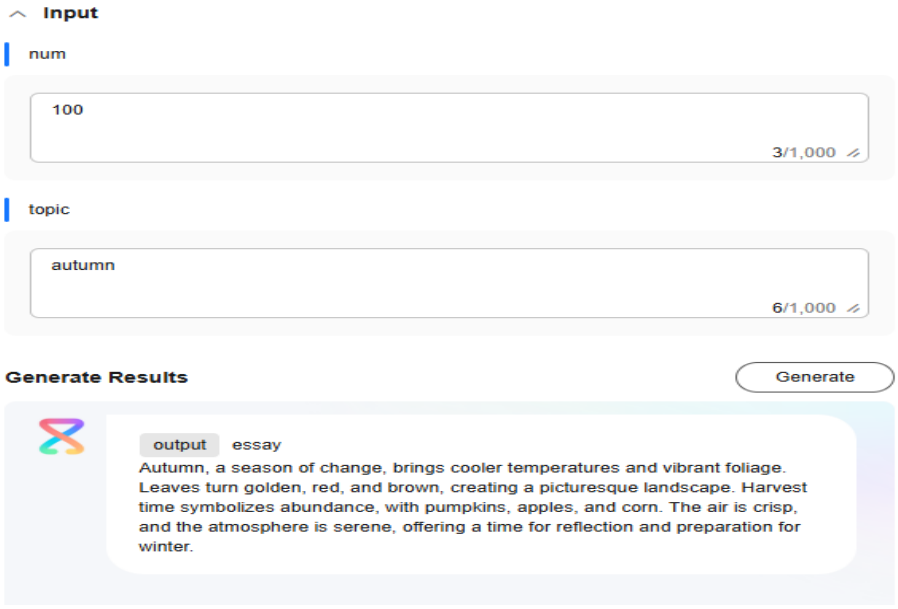
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-18 Configuring variables



6. Commission data instructions.
   - In the **Commissioning** > **Model** area, select the model required by the instruction and click **Configure Hyperparameters** to customize hyperparameter values.
   - In the **Commissioning** > **Input** area, you can view the effect by assigning values to variables.

   Figure 4-19 Instruction commissioning



7. After the commissioning is complete, click **Create Now** to create the data instruction.

   The created data instruction is displayed on the **Manage Synthetic Instructions** > **Custom** page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.5.2.3 Labeling Text Datasets

Creating a Text Dataset Labeling Task

Before labeling a text dataset, import data. For details, see **Importing Data to the Pangu Platform**.

The data labeling function allows you to create annotation tasks, label datasets (annotation tasks), review labeled datasets (review tasks), and manage annotation tasks (task management). The functions supported by different roles and the displayed frontend pages are slightly different. For details, see **Table 4-18**.

Table 4-18 Data labeling task permissions supported by different roles

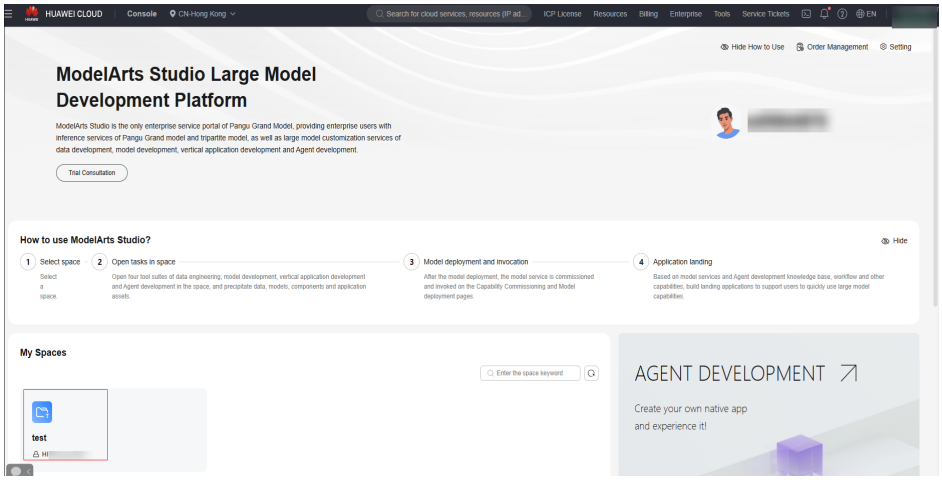
Role Name	Labeling Task Creation	Labeling Task	Labeling Review	Labeling Task Management
Super Admin	√	√	-	√
Administrator	√	√	-	√
Annotation administrator	√	√	-	√
Annotation operator	-	√	-	-
Annotation auditor	-	-	√	-

To create a text dataset annotation task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-20 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**. On the displayed page, click **Create annotation task** in the upper right corner.
3. On the **Create annotation task** page, select the text dataset to be annotated and select annotation items. The annotation items vary depending on the data file type. You can select annotation items as prompted.

   The single-turn Q&A annotation items support the AI-assisted annotation function. If this function is enabled, you need to select a deployed NLP service as the AI-assisted annotation model.
4. You can enable the **multiplayer job** function. After this function is enabled, you can select multiple persons to complete the operation. In addition, the review function can be enabled if required. Configure labeling assignment and review by referring to **Table 4-19**.

Table 4-19 Labeling assignment and review configuration

Category	Parameter Name	Description
Labeling assignment	Annotator	Add annotators and the number of annotations.
Labeling review	Is reviewed	• **No**: The review operation is not performed after labeling. • **Yes**: The reviewer checks the annotation content of the annotator. If any problem is found, the reviewer can specify the reason and reject the annotation data. The annotator needs to label the data again.
	Reviewer	Add reviewers and the number of reviewers.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
	Review requirements	• Full review: The reviewer needs to manually review all data records one by one. • Partial review: If the labeling quality of some data is high, the reviewer can submit the remaining data for review in one-click mode. By default, the data is approved and the review task is complete.

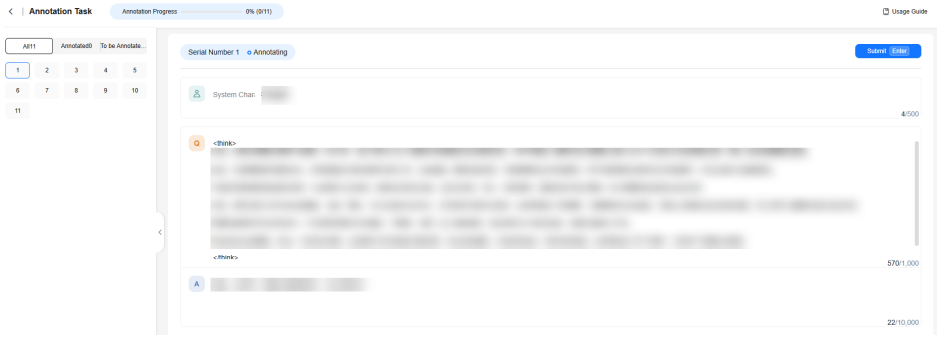
5. After the configuration is complete, click **Complete creation**.
6. On the **Annotation Task** page, click **Work** in the **Operation** column of the current labeling task to execute the task.

   For the "Annotate operators" role, you can click **Annotation** to execute the job.

   If you need to transfer the labeling task to other personnel, click **Transfer**, set the transfer personnel and quantity, and click **OK**.
7. On the labeling page, label data record one by one.

   As shown in **Figure 4-21**, if the data of a single-turn Q&A is labeled, you need to check whether the question (Q) and answer (A) are correct one by one. If the question or answer is incorrect, you can edit the question or answer again.

Figure 4-21 Example of labeling a text dataset



8. After labeling a piece of data, click **Submit** to label the remaining data. After all data is labeled, a message is displayed, indicating that the labeling task is successful.

NOTE

After data labeling is complete, if you do not need to review the labeling result, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

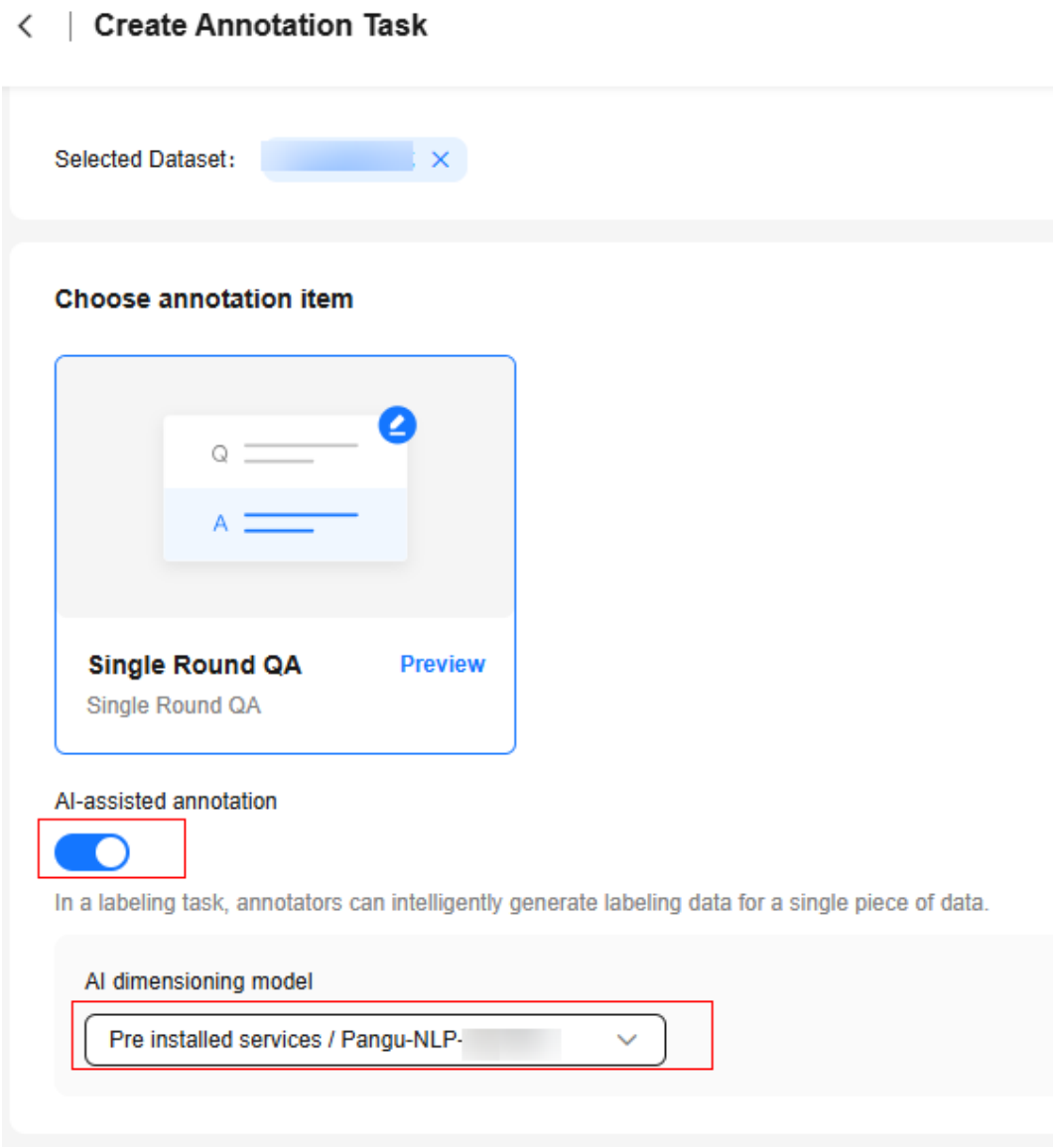
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

AI-assisted Labeling

In a labeling task, the labeler can use intelligently generated label for a single data record. The procedure is as follows:

Step 1 On the **Create Annotation Task** page, select a dataset, enable **AI-assisted annotation**, and click **Complete creation**.

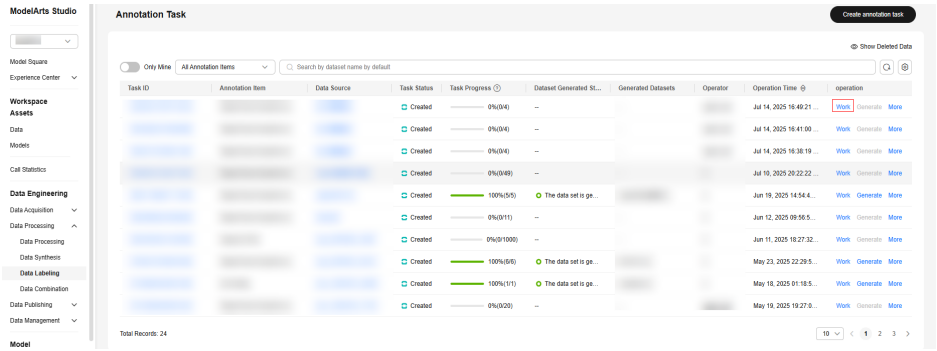
Figure 4-22 AI-assisted labeling



Step 2 On the **Annotation Task** page, click **Work** in the **Operation** column of the task to access the labeling page.

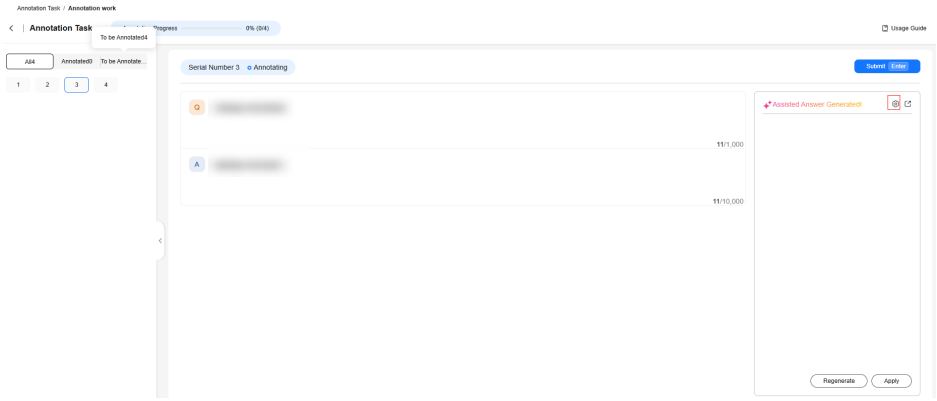
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-23 Labeling tasks



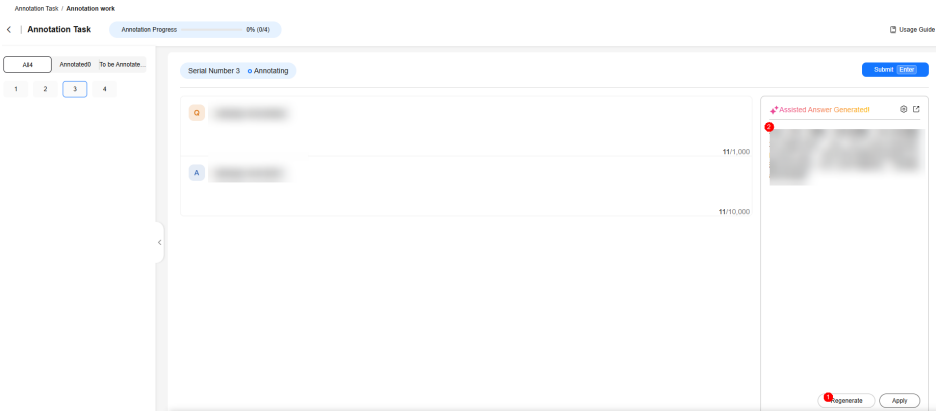
Step 3 Click the **Settings** button, modify the AI labeling model, and click **OK**.

Figure 4-24 Settings



Step 4 Click **Regenerate** to generate a new answer.

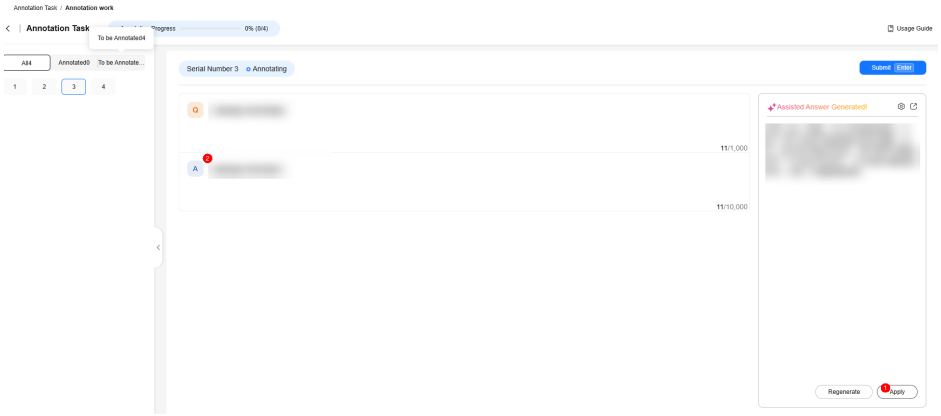
Figure 4-25 Regenerate



Step 5 Click **Apply**. The regenerated content is displayed in the answer box.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-26 Apply



----End

Reviewing a Labeled Text Dataset

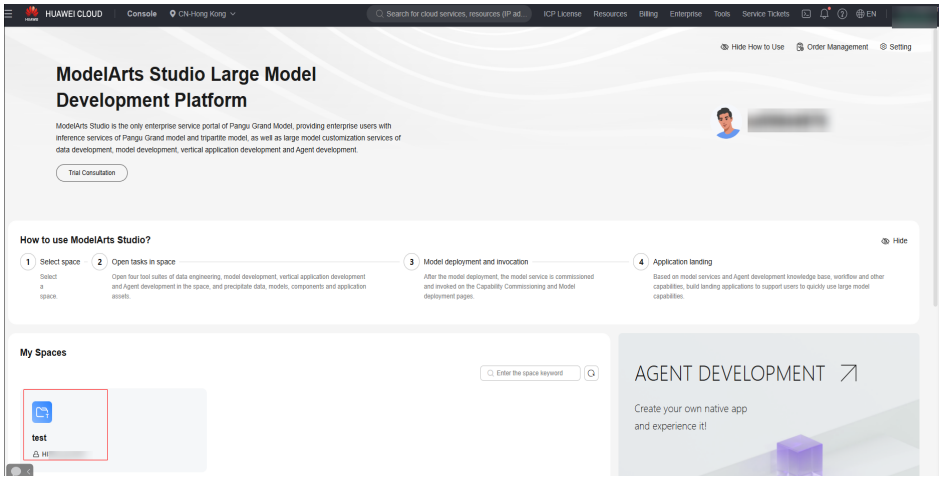
If the labeling review function is enabled in **Creating a Text Dataset Labeling Task**, the labeling result can be reviewed after the labeling is complete. If the data fails to pass the review, you can enter the failure cause and reject the data to the annotator for re-labeling.

You must have the permission of the **Annotate auditors** role.

To review the labeling result of a text dataset, perform the following steps:

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

   Figure 4-27 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**.
3. Click **Review** to go to the review page and review the data.

   If you want to transfer the review task to another person, click **Transfer**, set the transferee and transfer quantity, and click **OK**.
4. On the review page, you can click **Approve** or **Reject** to review the data record one by one until all data is reviewed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

NOTE

After the data labeling review is complete, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

Managing Labeled Text Datasets

The platform allows super administrators, administrators, and labeling administrators to perform the following operations on labeled datasets:

- Generate: After data labeling is reviewed, the super administrator, administrator, or labeling administrator needs to click **Generate** in the **Operation** column on the **Annotation Task** page to generate a processed dataset.

  To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.
- Stop: The administrator can click **More** > **Stop** in the **Operation** column on the **Annotation Task** page to stop the current labeling task.
- Edit: If the labeling job supports the AI-assisted annotation and review functions, the administrator can click **More** > **Edit** in the **Operation** column on the **Annotation Task** page, determine whether to enable the AI-assisted annotation function, and set review requirements.
- Delete: The administrator can click **More** > **Delete** in the **Operation** column on the **Annotation Task** page to delete the current annotation task.

4.5.2.4 Combining Text Datasets Based on a Specific Ratio

Data combination is a process of combining multiple datasets based on a specific ratio and publishing the combined dataset. A proper ratio ensures the diversity, balance, and representativeness of datasets.

If a single dataset meets your requirements, skip this section and proceed with **Publishing Text Datasets**.

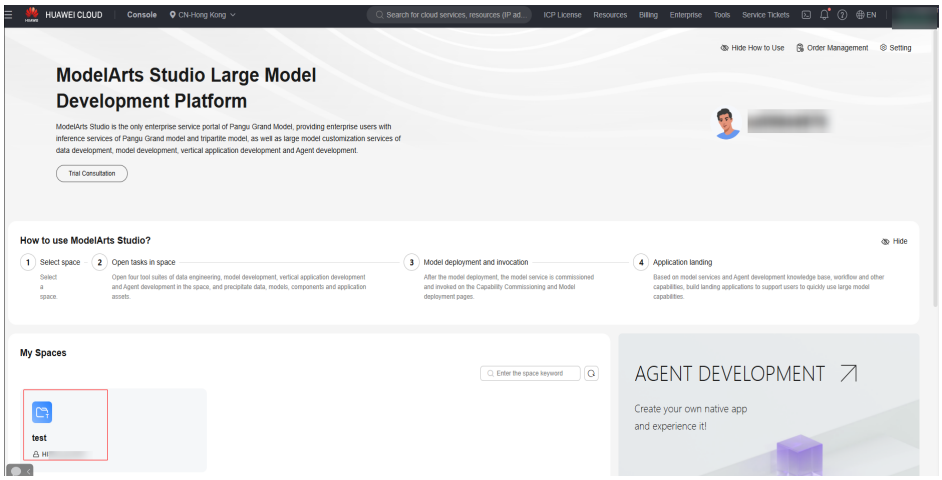
Creating a Text Data Combination Task

To create a text dataset combination task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

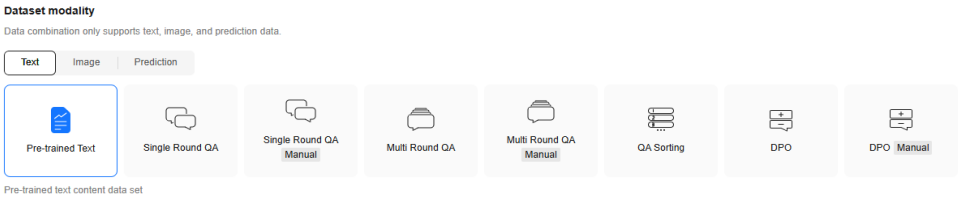
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-28 My Spaces



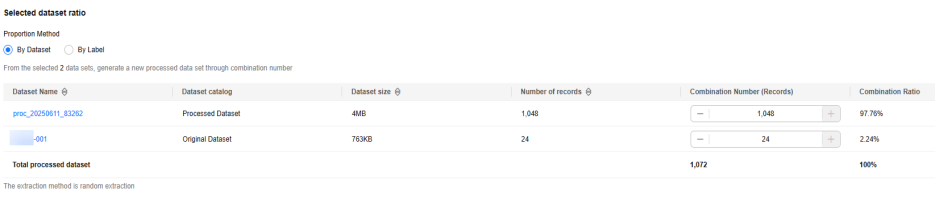
2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Combine Task**. On the displayed page, click **Create data combine** in the upper right corner.
3. In the **Dataset Modality** area, select the dataset modality for which data combination is to be performed. Text, image, and prediction datasets can be combined, as shown in **Figure 4-29**.

Figure 4-29 Dataset modality



4. In the **Select Dataset** area, select at least two image datasets and click **Next**.
5. In the data combination area, two modes are supported: by dataset and by label.
   - By Dataset: You can set the number of data records in the datasets to be combined, as shown in **Figure 4-30**.

     Figure 4-30 Example of setting the number of data records in the datasets to be combined

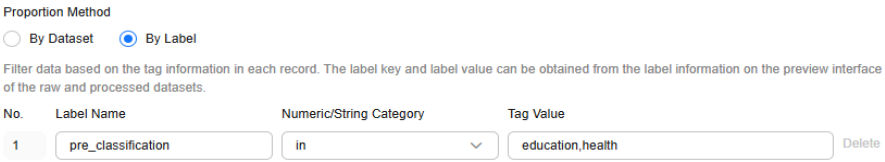
     

   - By label: This mode is applicable to text datasets processed by the data labeling operator. You can obtain the label name and value on the dataset details page after performing the operations in **Processing Text Datasets**.

     Figure 4-31 shows an example.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-31 Example of setting the labels for filtering the data to be combined



6. After the data combination configuration is complete, click **Next** in the lower right corner to go to the resource configuration page and select whether to automatically generate a processed dataset.

   Resource Allocation:

   Click ^ to expand resource configuration and set task resources. You can also customize parameters. Click **Add Parameters** and enter the parameter name and value.

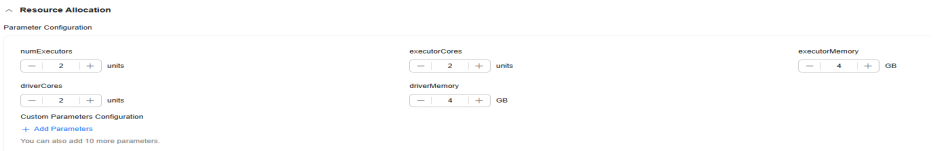
Table 4-20 Parameter configuration

Parameter Name	Description
numExecutors	Number of executors. The default value is 2. An executor is a process running on a worker node. It executes tasks and returns the calculation result to the driver. One core in an executor can run one task at the same time. Therefore, more tasks can be processed at the same time if you increase the number of the executors. You can add executors (if they are available) to process more tasks concurrently and improve efficiency. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16.
executorCores	Number of CPU kernels used by each executor process. The default value is 2. Multiple cores in an executor can run multiple tasks at the same time, which increases the task concurrency. However, because all cores share the memory of an executor, you need to balance the memory and the number of cores. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
executorMemory	Memory size used by each Executor process. The default value is 4. The executor memory is used for task execution and communication. You can increase the memory for a job that requires a great number of resources, and run small jobs concurrently with a smaller memory. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Description
driverCores	Number of CPU kernels used by each driver process. The default value is 2. The driver schedules jobs and communicates with executors. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.
driverMemory	Memory used by the driver process. The default value is 4. The driver schedules jobs and communicates with executors. Add driver memory when the number and parallelism level of the tasks increases. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.

Figure 4-32 Resource Allocation



Automatically Generate Processing Dataset

Select this option and configure the information for generating a processed dataset, as shown in **Figure 4-33**. Click **Confirm** in the lower right corner. The platform starts the data combination task. After the combination task is successfully executed, a processed dataset is automatically generated.

If you do not select this option, click **OK** in the lower right corner. The platform starts the data combination task. After the combination task is successfully executed, manually generate a processed dataset.

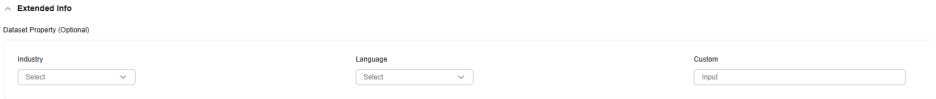
Figure 4-33 Automatically Generate Processing Dataset



(Optional) Extended Info:

You can select the industry and language, or customize dataset properties.

Figure 4-34 Extended Info



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

7. Click **Confirm**. On the **Data Combine Task** page, after the task is executed successfully, check that the status is **Success**.
8. Click **Generate** in the **Operation** column to generate a published dataset.

   To view the published dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Published Dataset** tab.

4.5.3 Processing Image Datasets

4.5.3.1 Processing Image Datasets

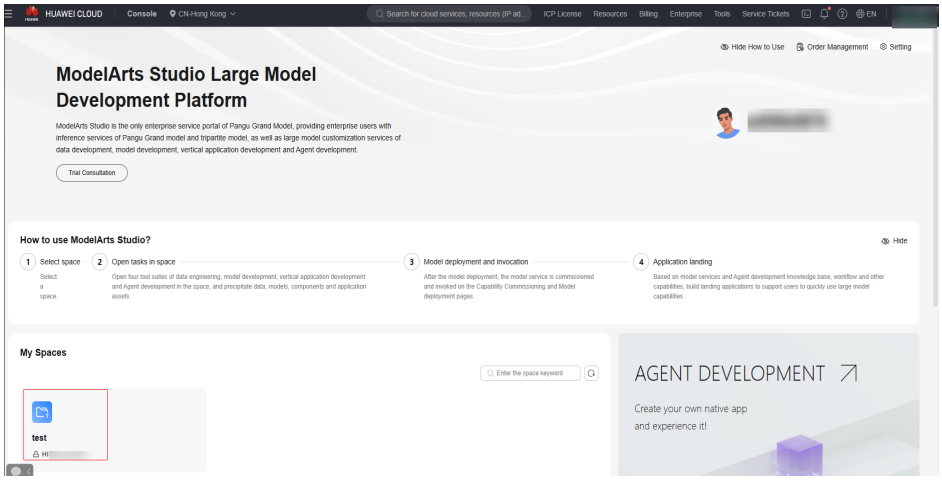
Before processing an image dataset, import data. For details, see **Importing Data to the Pangu Platform**.

Creating an Image Dataset Processing Task

To create an image dataset processing task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-35 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Processing Tasks**. Click **Create Processing Job** in the upper right corner.
3. On the **Create Processing Job** page, select the image dataset to be processed and click **Next**.
4. Go to the **processing step arrangement** page. For details about available processing operators for image datasets, see **Table 4-31**.
   a. In the **Adding Operator** pane on the left, select the required operators. You can select one or more operators based on the actual scenario.

      The platform supports preset operators and custom operators. For details about how to create a custom operator, see **Custom Data Processing Operators**.
   b. On the **processing step orchestration** page on the right, set operator parameters. You can drag ⠿ on the right to adjust the operator execution sequence.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

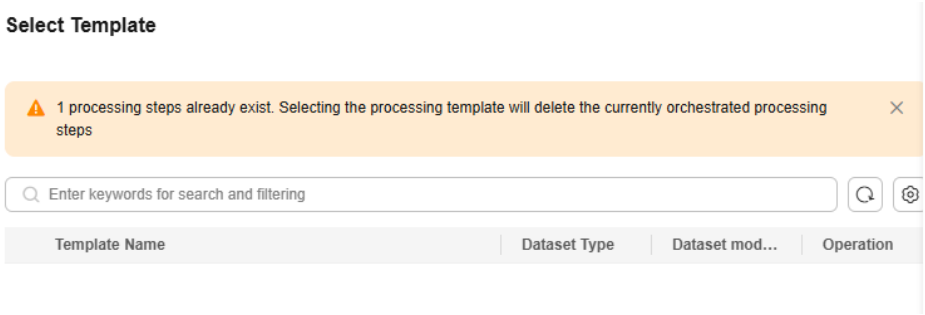
Figure 4-36 Operator orchestration



c. During orchestration, you can click **Save new template** in the upper right corner to save the current orchestration process as a template. During the creation of subsequent data processing tasks, you can select a processing template.

   If you select a processing template, the orchestrated processing steps will be deleted.

Figure 4-37 Selecting a processing template



5. After the processing steps are orchestrated, click **Next** to go to the **Task Configuration** page.
   - Resource Allocation

     Click ⌄ to expand resource configuration and set task resources. You can also customize parameters. Click **Add Parameters** and enter the parameter name and value.

Figure 4-38 Resource Allocation

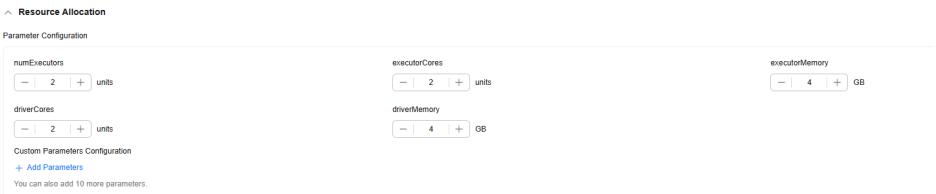


Table 4-21 describes the parameters.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-21 Parameter configuration

Parameter Name	Description
numExecutors	Number of executors. The default value is 2. An executor is a process running on a worker node. It executes tasks and returns the calculation result to the driver. One core in an executor can run one task at the same time. Therefore, more tasks can be processed at the same time if you increase the number of the executors. You can add executors (if they are available) to process more tasks concurrently and improve efficiency. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16.
executorCores	Number of CPU kernels used by each executor process. The default value is 2. Multiple cores in an executor can run multiple tasks at the same time, which increases the task concurrency. However, because all cores share the memory of an executor, you need to balance the memory and the number of cores. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
executorMemory	Memory size used by each Executor process. The default value is 4. The executor memory is used for task execution and communication. You can increase the memory for a job that requires a great number of resources, and run small jobs concurrently with a smaller memory. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
driverCores	Number of CPU kernels used by each driver process. The default value is 2. The driver schedules jobs and communicates with executors. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.
driverMemory	Memory used by the driver process. The default value is 4. The driver schedules jobs and communicates with executors. Add driver memory when the number and parallelism level of the tasks increases. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Automatically Generate Processing Dataset

  Select and configure the information about the generated dataset, as shown in **Figure 4-39**. Click **OK** in the lower right corner. The platform starts the data combination task. After the task is successfully executed, a processed dataset is automatically generated.

  If you do not select this option, click **OK** in the lower right corner. The platform starts the combination task. After the combination task is successfully executed, manually generate a processed dataset.

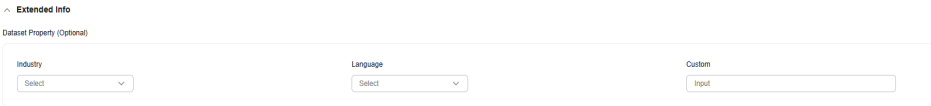
  Figure 4-39 Automatically Generate Processing Dataset

- (Optional) Extended Info

  You can select the industry and language, or customize dataset properties.

  Figure 4-40 Extended Info

6. Click **Start Process**. After the data processing job is successfully executed, the status changes from **Running** to **Succeeded**, indicating that the data has been processed.

   NOTE

   - After data processing is complete, if you do not need to use the data labeling and data synthesis functions, click **Generate** in the **Operation** column on the **Processing Tasks** page to generate a processed dataset.
   - To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

4.5.3.2 Labeling Image Datasets

Creating an Image Dataset Labeling Task

Before labeling an image dataset, import data. For details, see **Importing Data to the Pangu Platform**.

The data labeling function allows you to create annotation tasks, label datasets (annotation tasks), review labeled datasets (review tasks), and manage annotation tasks (task management). The functions supported by different roles and the displayed frontend pages are slightly different. For details, see **Table 4-22**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

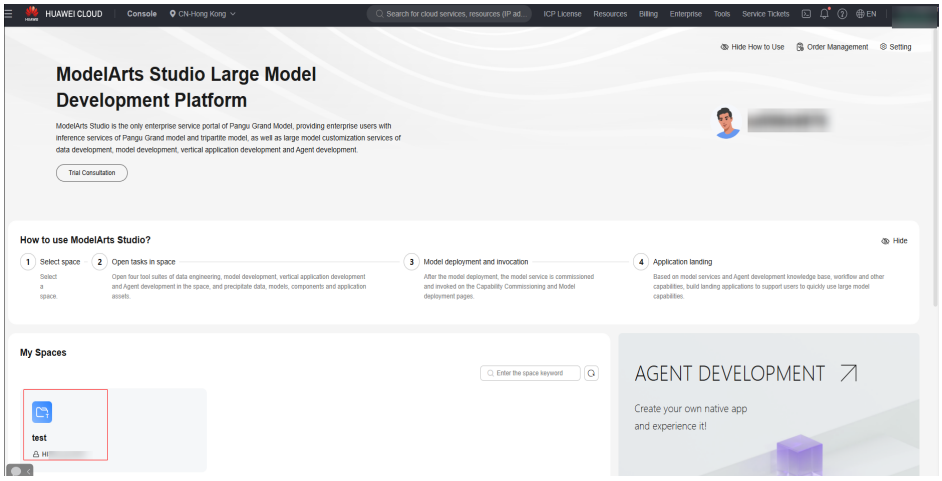
Table 4-22 Data labeling task permissions supported by different roles

Role Name	Labeling Task Creation	Labeling Task	Labeling Review	Labeling Task Management
Super Admin	√	√	-	√
Administrator	√	√	-	√
Annotation administrator	√	√	-	√
Annotation operator	-	√	-	-
Annotation auditor	-	-	√	-

To create an image dataset labeling task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 4-41 My Spaces



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**. On the displayed page, click **Create annotation task** in the upper right corner.
3. On the **Create annotation task** page, select the image dataset to be labeled and select labeling items.

   If you select **Image Caption**, you can enable the AI pre-annotation function. AI pre-annotation automatically generates labeling content and does not overwrite the original dataset. The annotation content can be used as a reference for annotation personnel to improve annotation efficiency.
4. You can enable the **multiplayer job** function. After this function is enabled, you can select multiple persons to complete the operation. In addition, the review function can be enabled if required. Configure labeling assignment and review by referring to **Table 4-23**.

Table 4-23 Labeling assignment and review configuration

Category	Parameter Name	Description
Labeling assignment	Annotator	Add annotators and the number of annotations.
	Annotation requirements	If you select **Image Caption** and enable AI pre-annotation, you can set the annotation requirements in either of the following ways: • **Full annotation**: The labeling personnel need to manually label all data before submitting the annotations. • **Partial annotation**: After confirming that the AI pre-annotation meets the requirements, you can directly use the AI pre-annotation function to label the dataset and submit the annotations.
Labeling review	Is reviewed	• **No**: The review operation is not performed after labeling. • **Yes**: The reviewer checks the annotation content of the annotator. If any problem is found, the reviewer can specify the reason and reject the annotation data. The annotator needs to label the data again.
	Reviewer	Add reviewers and the number of reviewers.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
	Review requirements	• Full review: The reviewer needs to manually review all data records one by one. • Partial review: If the labeling quality of some data is high, the reviewer can submit the remaining data for review in one-click mode. By default, the data is approved and the review task is complete.

5. After the configuration is complete, click **Complete creation**.
6. On the **Annotation Task** page, click the current labeling task to execute the task.

   For the "Annotate operators" role, you can click **Annotation** to execute the job.

   If you need to transfer the labeling task to other personnel, click **Transfer**, set the transfer personnel and quantity, and click **OK**.
7. On the labeling page, label data record one by one.
8. After labeling a piece of data, click **Submit** to label the remaining data. After all data is labeled, a message is displayed, indicating that the labeling task is successful.

   If AI pre-annotation and partial annotation are configured during labeling task creation, you can click **Submit All Labeled Data** in the upper right corner after annotating some data so that the AI model can automatically label the remaining data.

NOTE

After data labeling is complete, if you do not need to review the labeling result, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

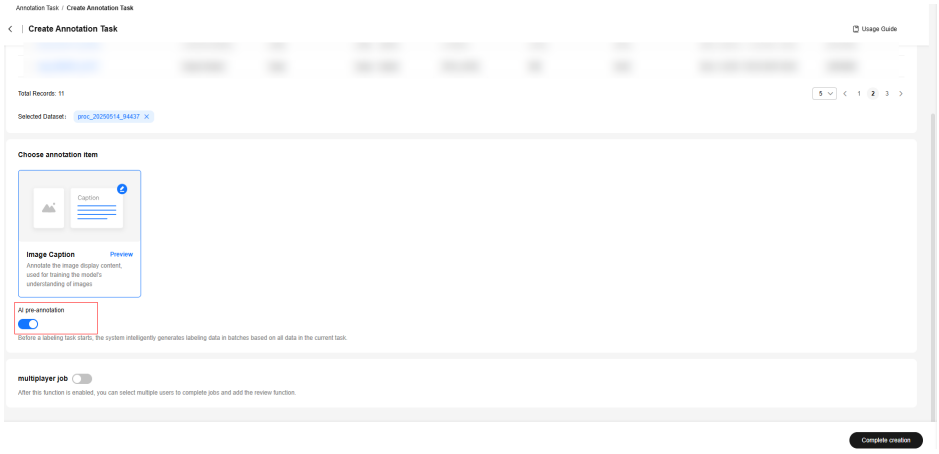
AI Pre-labeling

Before starting a labeling task, the system intelligently generates labeled data for all data in the task in batches. The procedure is as follows:

Step 1 On the **Create Annotation Task** page, select a dataset, enable **AI pre-annotation**, and click **Complete creation**.

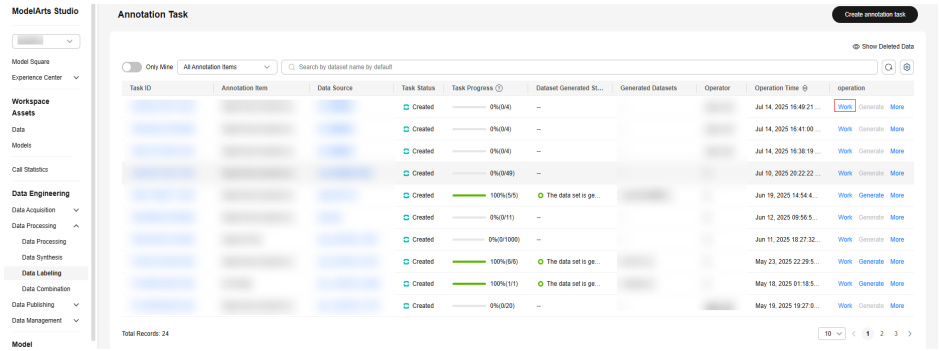
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-42 AI pre-labeling



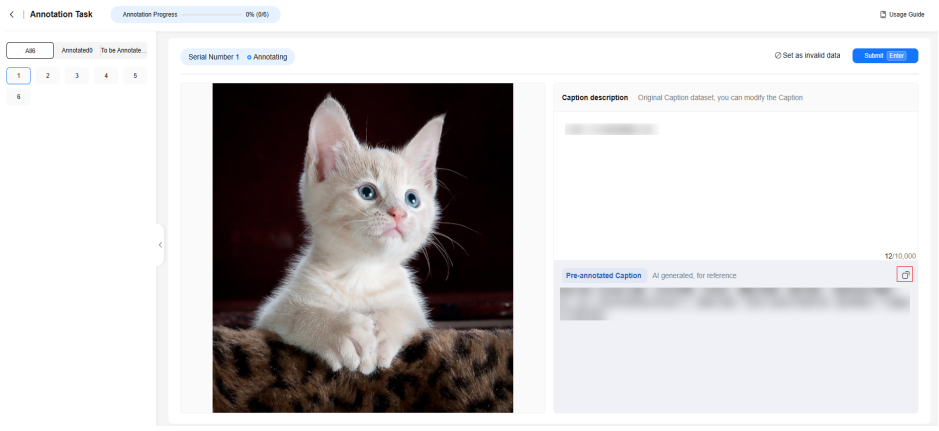
Step 2 On the **Annotation Task** page, click **Work** in the **Operation** column of the task to access the labeling page.

Figure 4-43 Labeling tasks



Step 3 Copy the content in the **Pre-annotated Caption** box and paste it to the **Caption description** box above. You can modify the caption and click **Submit** in the upper right corner.

Figure 4-44 Pre-labeled caption



----End

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Reviewing a Labeled Image Dataset

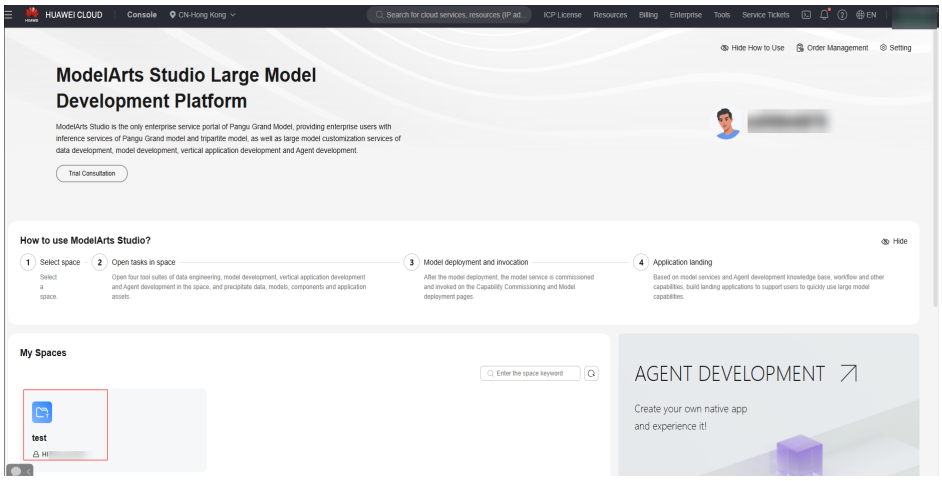
If the labeling review function is enabled in **Creating an Image Dataset Labeling Task**, the labeling result can be reviewed after the labeling is complete. If the data fails to pass the review, you can enter the failure cause and reject the data to the annotator for re-labeling.

You must have the permission of the **Annotate auditors** role.

To review the labeling result of an image dataset, perform the following steps:

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

   Figure 4-45 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**.
3. Click **Review** to go to the review page and review the data.

   If you want to transfer the review task to another person, click **Transfer**, set the transferee and transfer quantity, and click **OK**.
4. On the review page, you can click **Approve** or **Reject** to review the data record one by one until all data is reviewed.

NOTE

After the data labeling review is complete, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

Managing Labeled Image Datasets

The platform allows super administrators, administrators, and labeling administrators to perform the following operations on labeled datasets:

- Generate: After data labeling is reviewed, the super administrator, administrator, or labeling administrator needs to click **Generate** in the **Operation** column on the **Annotation Task** page to generate a processed dataset.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

- Stop: The administrator can click **More** > **Stop** in the **Operation** column on the **Annotation Task** page to stop the current labeling task.
- Edit: If the labeling job supports the AI-assisted annotation and review functions, the administrator can click **More** > **Edit** in the **Operation** column on the **Annotation Task** page, determine whether to enable the AI-assisted annotation function, and set review requirements.
- Delete: The administrator can click **More** > **Delete** in the **Operation** column on the **Annotation Task** page to delete the current annotation task.

4.5.3.3 Combining Image Datasets Based on a Specific Ratio

Data combination is a process of combining multiple datasets based on a specific ratio and publishing the combined dataset. A proper ratio ensures the diversity, balance, and representativeness of datasets.

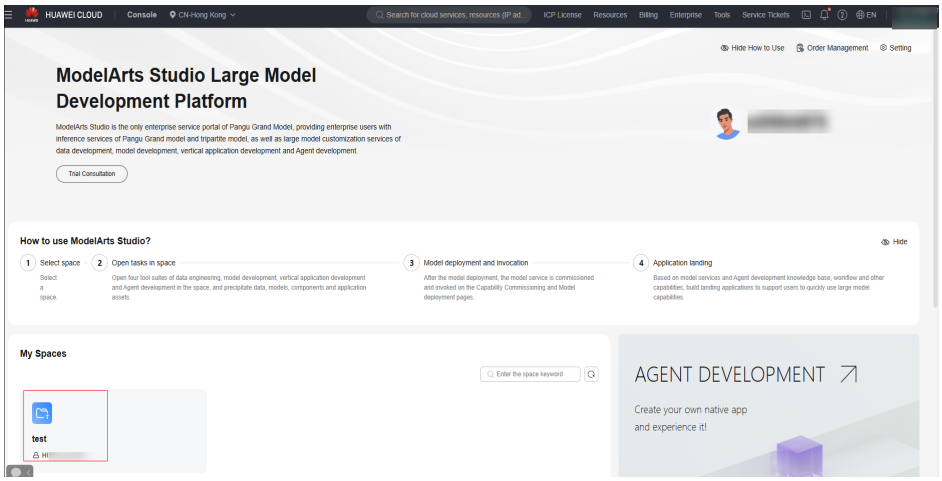
If a single dataset meets your requirements, skip this section and proceed with **Publishing Image Datasets**.

Creating an Image Data Combination Task

To create an image dataset combination task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-46 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Combine Task**. On the displayed page, click **Create data combine** in the upper right corner.
3. In the **Select Dataset** area, select at least two image datasets and click **Next**.
4. On the **Data Combine** page, set the ratio of different datasets and click **Next**.
5. After the data combination configuration is complete, click **Next** in the lower right corner to go to the resource configuration page and select whether to automatically generate a processed dataset.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

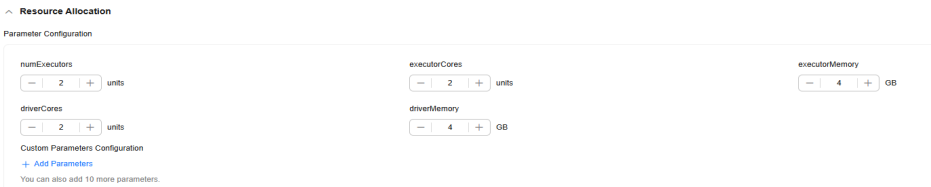
- Resource Allocation

  Click ^ to expand resource configuration and set task resources. You can also customize parameters. Click **Add Parameters** and enter the parameter name and value.

  Table 4-24 Parameter configuration

Parameter Name	Description
numExecutors	Number of executors. The default value is 2. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16.
executorCores	Number of CPU kernels used by each executor process. The default value is 2. numExecutors x executorMemory must be greater than or equal to 4 and less than or equal to 16. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
executorMemory	Memory size used by each Executor process. The default value is 4. The ratio of executorCores to executorMemory must be in the range of 1:2 to 1:4.
driverCores	Number of CPU kernels used by each driver process. The default value is 2. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.
driverMemory	Memory used by the driver process. The default value is 4. The ratio of driverCores to driverMemory must be in the range of 1:2 to 1:4.

  Figure 4-47 Resource Allocation



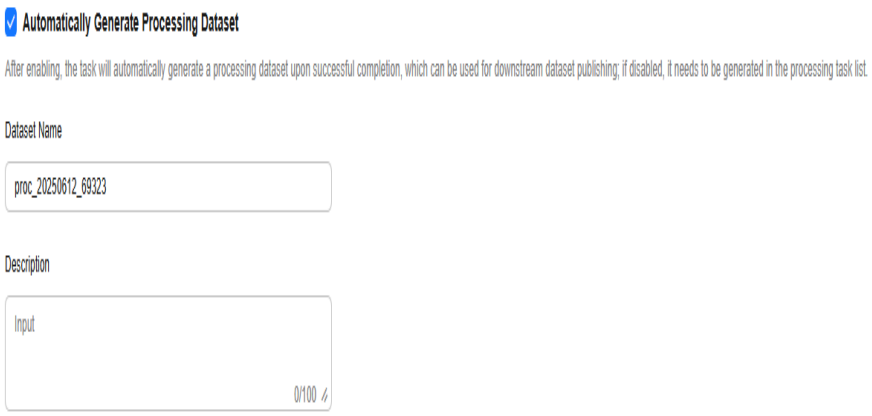
- Automatically Generate Processing Dataset

  Select and configure the information about the generated dataset, as shown in **Figure 4-48**. Click **OK** in the lower right corner. The platform starts the data combination task. After the task is successfully executed, a processed dataset is automatically generated.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

If you do not select this option, click **OK** in the lower right corner. The platform starts the combination task. After the combination task is successfully executed, manually generate a processed dataset.

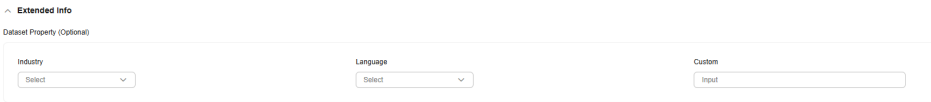
Figure 4-48 Automatically Generate Processing Dataset



- (Optional) Extended Info

  You can select the industry and language, or customize dataset properties.

  Figure 4-49 Extended Info

6. Click **OK**. On the **Data Combine Task** page, after the task is executed successfully, check that the status is **Success**.
7. Click **Generate** in the **Operation** column to generate a published dataset.

   To view the published dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Published Dataset** tab.

4.5.4 Processing Video Datasets

4.5.4.1 Processing Video Datasets

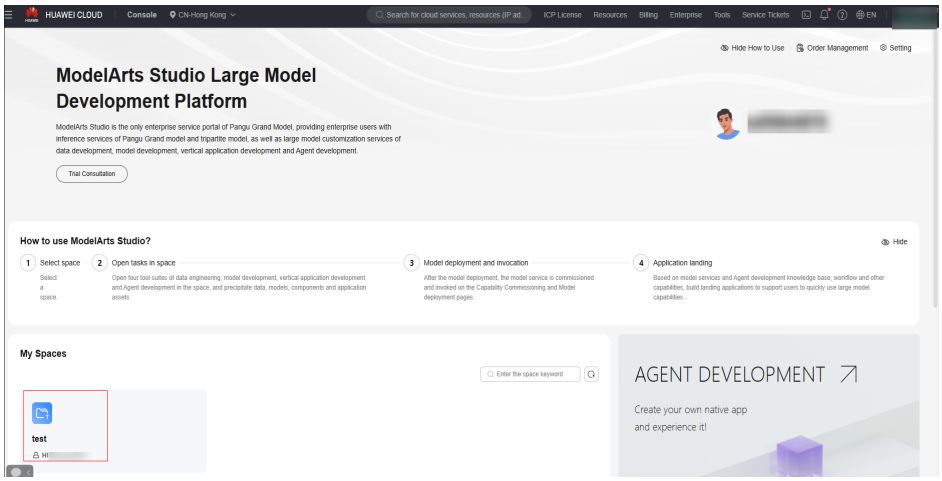
Before processing a video dataset, import data. For details, see **Importing Data to the Pangu Platform**.

Creating a Video Dataset Processing Task

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-50 My Spaces



2. In the navigation pane, choose **Data engineering** > **Data Processing** > **Processing Tasks**. Click **Create Processing Job** in the upper right corner.
3. On the **Create Processing Job** page, select the video dataset to be processed and click **Next**.
4. Go to the **processing step arrangement** page. For details about available processing operators for video datasets, see **Table 4-30**.

   a. In the **Adding Operator** pane on the left, select the required operators. You can select one or more operators based on the actual scenario.

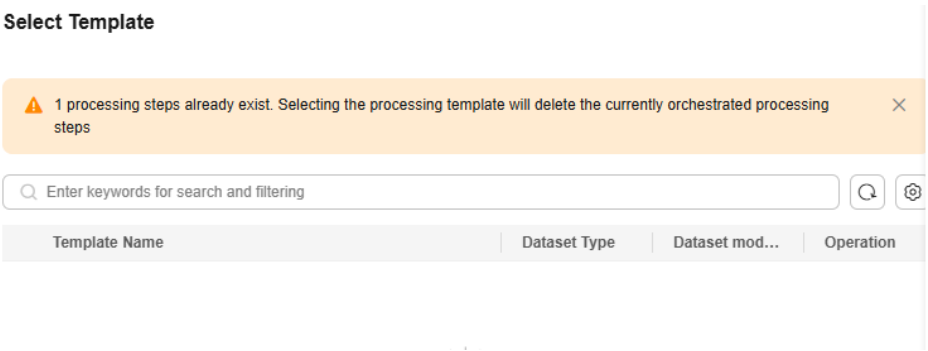
      The platform supports preset operators and custom operators. For details about how to create a custom operator, see **Custom Data Processing Operators**.

   b. On the **processing step orchestration** page on the right, set operator parameters. You can drag ⠿ on the right to adjust the operator execution sequence.

   c. During orchestration, you can click **Save new template** in the upper right corner to save the current orchestration process as a template. During the creation of subsequent data processing tasks, you can select a processing template.

      If you select a processing template, the orchestrated processing steps will be deleted.

Figure 4-51 Selecting a processing template



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

5. After the processing steps are orchestrated, click **Next** to go to the **Task Configuration** page.
   - Automatically Generate Processing Dataset

     Select and configure the information about the generated dataset, as shown in **Figure 4-52**. Click **OK** in the lower right corner. The platform starts the data combination task. After the task is successfully executed, a processed dataset is automatically generated.

     If you do not select this option, click **OK** in the lower right corner. The platform starts the combination task. After the combination task is successfully executed, manually generate a processed dataset.

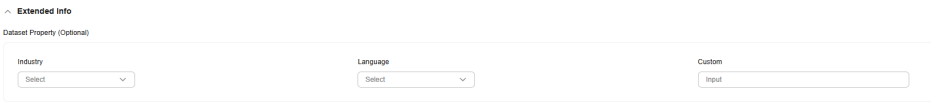
     Figure 4-52 Automatically Generate Processing Dataset

   - (Optional) Extended Info

     You can select the industry and language, or customize dataset properties.

     Figure 4-53 Extended Info

6. Click **Start Process**. After the data processing job is successfully executed, the status changes from **Running** to **Succeeded**, indicating that the data has been processed.

   NOTE

   - After data processing is complete, if you do not need to use the data labeling and data synthesis functions, click **Generate** in the **Operation** column on the **Processing Tasks** page to generate a processed dataset.
   - To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

4.5.4.2 Labeling Video Datasets

Creating a Video Dataset Labeling Task

Before labeling a video dataset, import data. For details, see **Importing Data to the Pangu Platform**.

The data labeling function allows you to create annotation tasks, label datasets (annotation tasks), review labeled datasets (review tasks), and manage annotation tasks (task management). The functions supported by different roles and the displayed frontend pages are slightly different. For details, see **Table 4-25**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

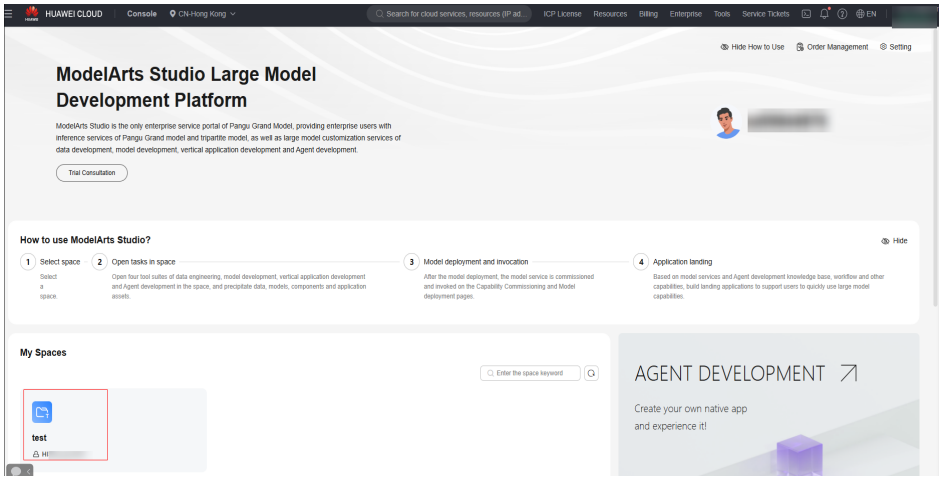
Table 4-25 Data labeling task permissions supported by different roles

Role Name	Labeling Task Creation	Labeling Task	Labeling Review	Labeling Task Management
Super Admin	√	√	-	√
Administrator	√	√	-	√
Annotation administrator	√	√	-	√
Annotation operator	-	√	-	-
Annotation auditor	-	-	√	-

To create a video dataset labeling task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 4-54 My Spaces



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**. On the displayed page, click **Create annotation task** in the upper right corner.
3. On the **Create annotation task** page, select the video dataset to be annotated and select annotation items, and click **Next**.

   If you select **Video Caption**, you can enable the AI pre-annotation function. AI pre-annotation automatically generates labeling content and does not overwrite the original dataset. The annotation content can be used as a reference for annotation personnel to improve annotation efficiency.
4. Enable the **multiplayer job** function if required. After this function is enabled, you can select multiple persons to complete the operation. In addition, the review function can be enabled if required. Configure labeling assignment and review by referring to **Table 4-26**.

   Table 4-26 Labeling assignment and review configuration

Category	Parameter Name	Description
Labeling assignment	Annotator	Add annotators and the number of annotations.
Labeling review	Is reviewed	• **No**: The review operation is not performed after labeling. • **Yes**: The reviewer checks the annotation content of the annotator. If any problem is found, the reviewer can specify the reason and reject the annotation data. The annotator needs to label the data again.
	Reviewer	Add reviewers and the number of reviewers.
	Review requirements	• Full review: The reviewer needs to manually review all data records one by one. • Partial review: If the labeling quality of some data is high, the reviewer can submit the remaining data for review in one-click mode. By default, the data is approved and the review task is complete.

5. After the configuration is complete, click **Complete creation**.
6. On the **Annotation Task** page, click the current labeling task to execute the task.

   For the "Annotate operators" role, you can click **Annotation** to execute the job.

   If you need to transfer the labeling task to other personnel, click **Transfer**, set the transfer personnel and quantity, and click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

7. On the labeling page, label data record one by one.
8. After labeling a piece of data, click **Submit** to label the remaining data. After all data is labeled, a message is displayed, indicating that the labeling task is successful.

   If AI pre-annotation and partial annotation are configured during labeling task creation, you can click **Submit All Labeled Data** in the upper right corner after annotating some data so that the AI model can automatically label the remaining data.

NOTE

After data labeling is complete, if you do not need to review the labeling result, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

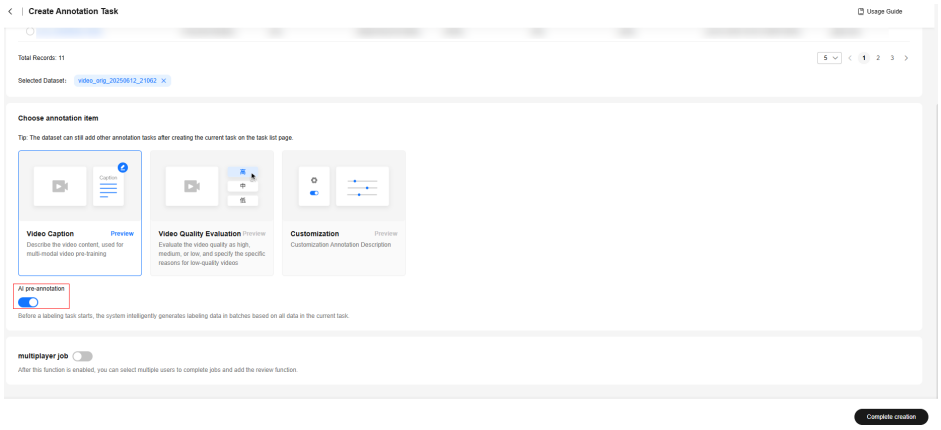
To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

AI Pre-labeling

Before starting a labeling task, the system intelligently generates labeled data for all data in the task in batches. The procedure is as follows:

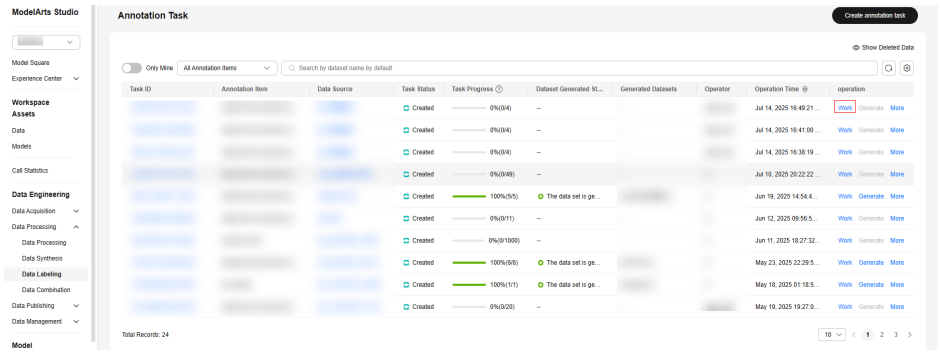
Step 1 On the **Create Annotation Task** page, select a dataset, enable **AI pre-annotation**, and click **Complete creation**.

Figure 4-55 AI pre-labeling



Step 2 On the **Annotation Task** page, click **Work** in the **Operation** column of the task to access the labeling page.

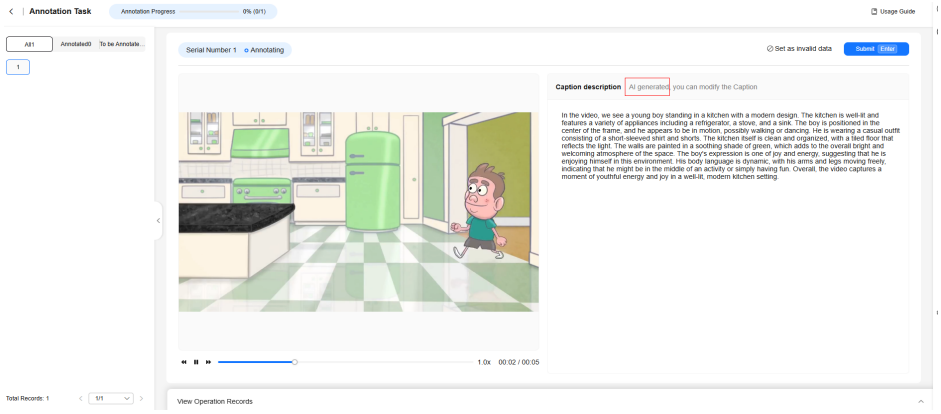
Figure 4-56 Labeling tasks



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Step 3 The **Caption description** box shows the caption generated by AI. You can modify the caption and click **Submit** in the upper right corner.

Figure 4-57 Caption description



----End

Reviewing a Labeled Video Dataset

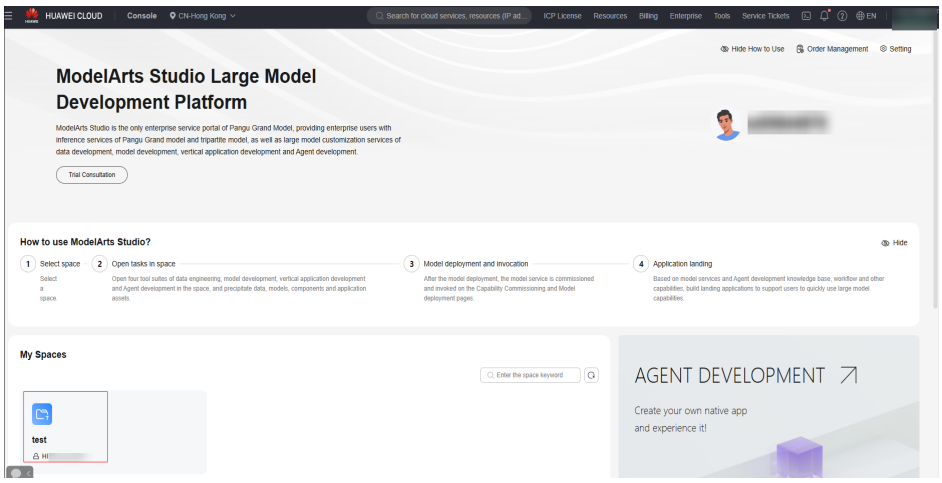
If the labeling review function is enabled in **Creating a Video Dataset Labeling Task**, the labeling result can be reviewed after the labeling is complete. If the data fails to pass the review, you can enter the failure cause and reject the data to the annotator for re-labeling.

You must have the permission of the **Annotate auditors** role.

To review the labeling result of a video dataset, perform the following steps:

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

   Figure 4-58 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**.
3. Click **Review** to go to the review page and review the data.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

If you want to transfer the review task to another person, click **Transfer**, set the transferee and transfer quantity, and click **OK**.

4. On the review page, you can click **Approve** or **Reject** to review the data record one by one until all data is reviewed.

NOTE

After the data labeling review is complete, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

Managing Labeled Video Datasets

The platform allows super administrators, administrators, and labeling administrators to perform the following operations on labeled datasets:

- Generate: After data labeling is reviewed, the super administrator, administrator, or labeling administrator needs to click **Generate** in the **Operation** column on the **Annotation Task** page to generate a processed dataset.

  To view the processed dataset, choose **Data Engineering** > **Data Management** > **Dataset**, and click the **Processed Dataset** tab.
- Stop: The administrator can click **More** > **Stop** in the **Operation** column on the **Annotation Task** page to stop the current labeling task.
- Edit: If the labeling job supports the AI-assisted annotation and review functions, the administrator can click **More** > **Edit** in the **Operation** column on the **Annotation Task** page, determine whether to enable the AI-assisted annotation function, and set review requirements.
- Delete: The administrator can click **More** > **Delete** in the **Operation** column on the **Annotation Task** page to delete the current annotation task.

4.5.5 Processing Weather Datasets

Before processing a weather dataset, import data. For details, see **Importing Data to the Pangu Platform**.

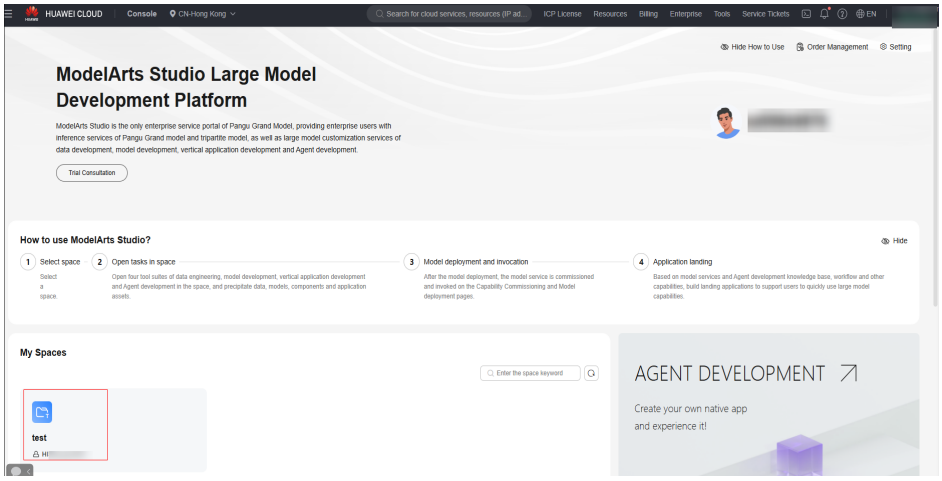
Creating a Weather Dataset Processing Task

To create a weather dataset processing task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-59 My Spaces



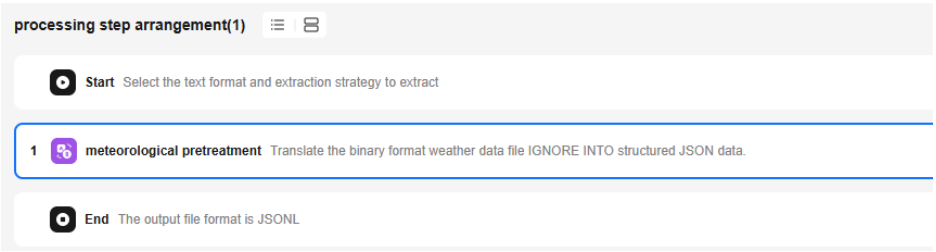
2. In the navigation pane, choose **Data engineering** > **Data Processing** > **Processing Tasks**. Click **Create Processing Job** in the upper right corner.
3. On the **Create Processing Job** page, select the weather dataset to be processed and click **Next**.
4. Go to the **processing step arrangement** page. For details about available processing operators for weather datasets, see **Table 4-32**.

   a. In the **Adding Operator** pane on the left, select the required operators. You can select one or more operators based on the actual scenario.

      The platform supports preset operators and custom operators. For details about how to create a custom operator, see **Custom Data Processing Operators**.

   b. On the **processing step orchestration** page on the right, set operator parameters. You can drag ⠿ on the right to adjust the operator execution sequence.

      Figure 4-60 Operator orchestration

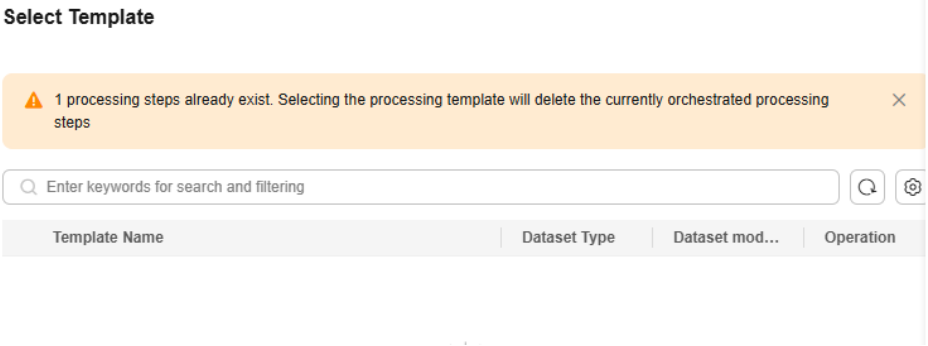
      

   c. During orchestration, you can click **Save new template** in the upper right corner to save the current orchestration process as a template. During the creation of subsequent data processing tasks, you can select a processing template.

      If you select a processing template, the orchestrated processing steps will be deleted.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-61 Selecting a processing template

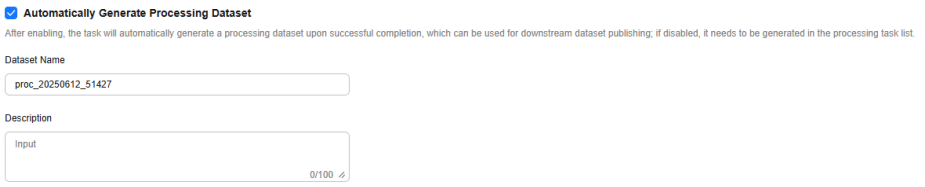


5. After the processing steps are orchestrated, click **Next** to go to the **Task Configuration** page.
   - Automatically Generate Processing Dataset

     Select and configure the information about the generated dataset, as shown in **Figure 4-62**. Click **OK** in the lower right corner. The platform starts the data combination task. After the task is successfully executed, a processed dataset is automatically generated.

     If you do not select this option, click **OK** in the lower right corner. The platform starts the combination task. After the combination task is successfully executed, manually generate a processed dataset.

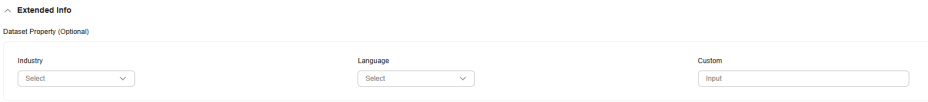
     Figure 4-62 Automatically Generate Processing Dataset

   - (Optional) Extended Info

     You can select the industry and language, or customize dataset properties.

     Figure 4-63 Extended Info

6. Click **Start Process**. After the data processing job is successfully executed, the status changes from **Running** to **Succeeded**, indicating that the data has been processed.

   NOTE

   - After data processing is complete, if you do not need to use the data labeling and data synthesis functions, click **Generate** in the **Operation** column on the **Processing Tasks** page to generate a processed dataset.
   - To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.5.6 Processing Audio Datasets

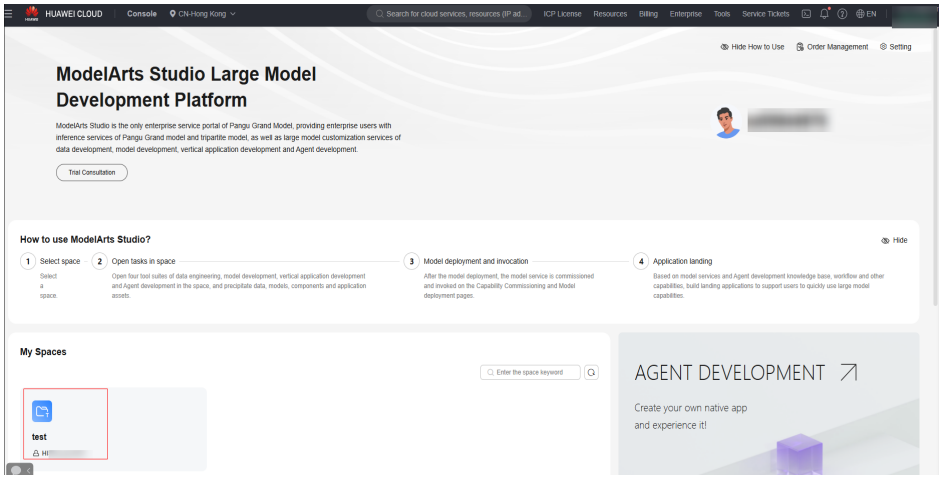
4.5.6.1 Processing Audio Datasets

Before processing an audio dataset, import data. For details, see **Importing Data to the Pangu Platform**.

Creating an Audio Dataset Processing Task

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-64 My Spaces

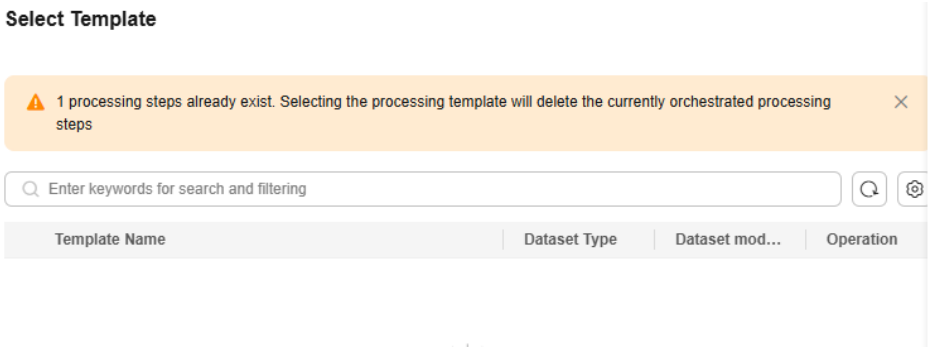
2. In the navigation pane, choose **Data engineering** > **Data Processing** > **Processing Tasks**. Click **Create Processing Job** in the upper right corner.
3. On the **Create Processing Job** page, select the audio dataset to be processed and click **Next**.
4. Go to the **processing step arrangement** page. For details about available processing operators for audio datasets, see **Audio Dataset Processing Operators**.
   a. In the **Adding Operator** pane on the left, select the required operators. You can select one or more operators based on the actual scenario.

      The platform supports preset operators and custom operators. For details about how to create a custom operator, see **Custom Data Processing Operators**.
   b. On the **processing step orchestration** page on the right, set operator parameters. You can drag ⠿ on the right to adjust the operator execution sequence.
   c. During orchestration, you can click **Save new template** in the upper right corner to save the current orchestration process as a template. During the creation of subsequent data processing tasks, you can select a processing template.

      If you select a processing template, the orchestrated processing steps will be deleted.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-65 Selecting a processing template



5. After the processing steps are orchestrated, click **Start Process** to start the processing task.

   After the data processing job is successfully executed, the status changes from **Running** to **Succeeded**, indicating that the data has been processed.

   NOTE

   After data processing is complete, if you do not need to use the data labeling function, click **Generate** in the **Operation** column on the **Processing Tasks** page to generate a processed dataset.

   To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

4.5.6.2 Labeling Audio Datasets

Creating an Audio Dataset Labeling Task

Before labeling an audio dataset, import data. For details, see **Importing Data to the Pangu Platform**.

The data labeling function allows you to create annotation tasks, label datasets (annotation tasks), review labeled datasets (review tasks), and manage annotation tasks (task management). The functions and GUIs vary depending on roles. For details, see **Table 4-27**.

Table 4-27 Data labeling task permissions supported by different roles

Role Name	Labeling Task Creation	Labeling Task	Labeling Review	Labeling Task Management
Super Admin	√	√	-	√
Administrator	√	√	-	√
Annotation administrator	√	√	-	√

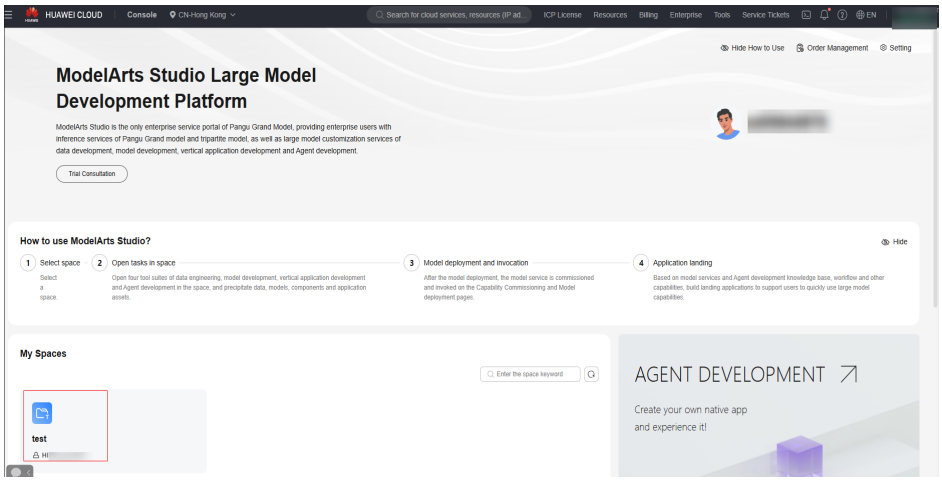
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Role Name	Labeling Task Creation	Labeling Task	Labeling Review	Labeling Task Management
Annotation operator	-	√	-	-
Annotation auditor	-	-	√	-

To create an audio dataset labeling task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-66 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**. On the displayed page, click **Create annotation task** in the upper right corner.
3. On the **Create annotation task** page, select the audio dataset to be annotated and select annotation items, and click **Next**.
4. Enable the **multiplayer job** function if required. After this function is enabled, you can select multiple persons to complete the operation. In addition, the review function can be enabled if required. Configure labeling assignment and review by referring to **Table 4-28**.

   Table 4-28 Labeling assignment and review configuration

Category	Parameter Name	Description
Labeling assignment	Annotator	Add annotators and the number of annotations.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
Labeling review	Is reviewed	• **No**: The review operation is not performed after labeling. • **Yes**: The reviewer checks the annotation content of the annotator. If any problem is found, the reviewer can specify the reason and reject the annotation data. The annotator needs to label the data again.
	Reviewer	Add reviewers and the number of reviewers.
	Review requirements	• Full review: The reviewer needs to manually review all data records one by one. • Partial review: If the labeling quality of some data is high, the reviewer can submit the remaining data for review in one-click mode. By default, the data is approved and the review task is complete.

5. After the configuration is complete, click **Complete creation**.
6. On the **Annotation Task** page, click the current labeling task to execute the task.

   For the "Annotate operators" role, you can click **Annotation** to execute the job.

   If you need to transfer the labeling task to other personnel, click **Transfer**, set the transfer personnel and quantity, and click **OK**.
7. On the labeling page, label data record one by one.
8. After labeling a piece of data, click **Submit** to label the remaining data. After all data is labeled, a message is displayed, indicating that the labeling task is successful.

NOTE

After data labeling is complete, if you do not need to review the labeling result, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

Reviewing a Labeled Audio Dataset

If the labeling review function is enabled in **Creating an Audio Dataset Labeling Task**, the labeling result can be reviewed after the labeling is complete. If the data fails to pass the review, you can enter the failure cause and reject the data to the annotator for re-labeling.

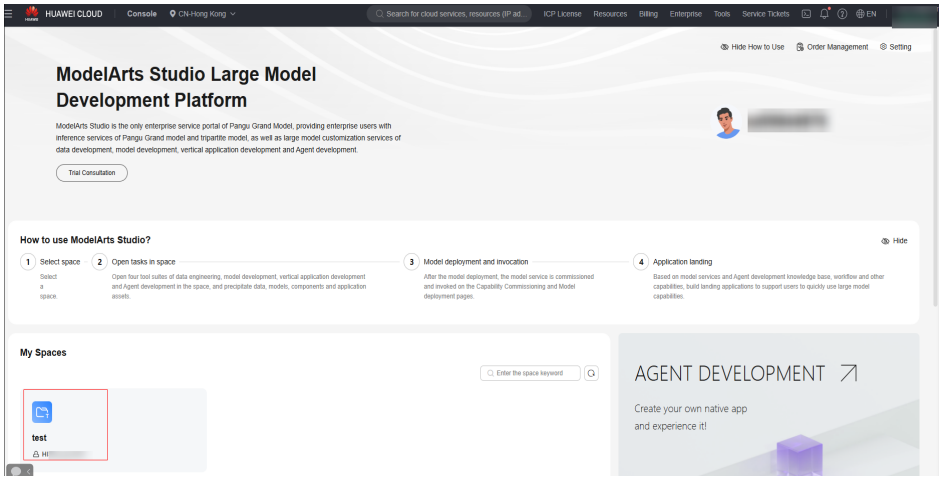
You must have the permission of the **Annotate auditors** role.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

To review the labeling result of a video dataset, perform the following steps:

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

   Figure 4-67 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Data Labeling**.
3. Click **Review** to go to the review page and review the data.

   If you want to transfer the review task to another person, click **Transfer**, set the transferee and transfer quantity, and click **OK**.
4. On the review page, you can click **Approve** or **Reject** to review the data record one by one until all data is reviewed.

NOTE

After the data labeling review is complete, click **Generate** on the **Annotation Task** > **Task Management** page to generate a processed dataset.

To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

Managing Labeled Audio Datasets

The platform allows super administrators, administrators, and labeling administrators to perform the following operations on labeled datasets:

- Generate: After data labeling is reviewed, the super administrator, administrator, or labeling administrator needs to click **Generate** in the **Operation** column on the **Annotation Task** page to generate a processed dataset.

  To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.
- Stop: The administrator can click **More** > **Stop** in the **Operation** column on the **Annotation Task** page to stop the current labeling task.
- Edit: If the labeling job supports the AI-assisted annotation and review functions, the administrator can click **More** > **Edit** in the **Operation** column on the **Annotation Task** page, determine whether to enable the AI-assisted annotation function, and set review requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Delete: The administrator can click **More** > **Delete** in the **Operation** column on the **Annotation Task** page to delete the current annotation task.

4.5.7 Processing Other Datasets

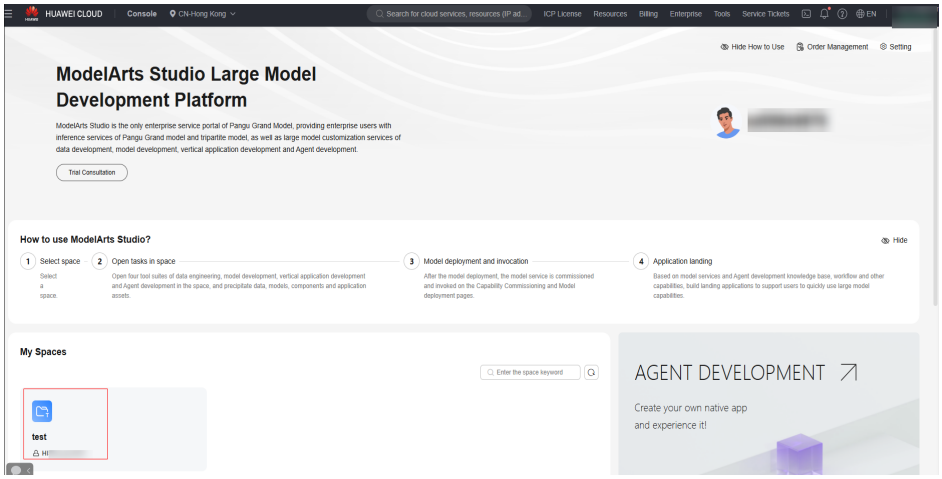
Before processing a dataset of other types, import data. For details, see **Importing Data to the Pangu Platform**.

Creating an Other-Type Dataset Processing Task

To create a processing task for a dataset of other types, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-68 My Spaces

2. In the navigation pane, choose **Data engineering** > **Data Processing** > **Processing Tasks**. Click **Create Processing Job** in the upper right corner.
3. On the **Create Processing Job** page, select the other type of dataset to be processed and click **Next**.
4. Go to the **processing step arrangement** page.
   a. In the **Adding Operator** pane on the left, select the required operators. You can select one or more operators based on the actual scenario.

      For other types of datasets, the platform supports only custom operators. For details about how to create a custom operator, see **Custom Data Processing Operators**.

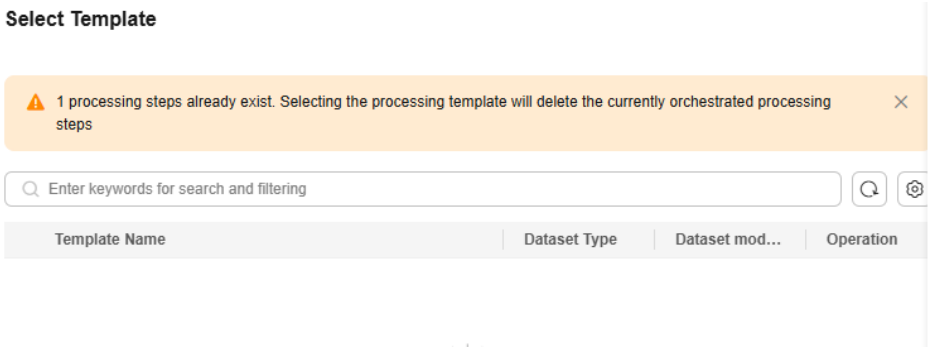
   b. On the **processing step orchestration** page on the right, set operator parameters. You can drag ⠿ on the right to adjust the operator execution sequence.

   c. During orchestration, you can click **Save new template** in the upper right corner to save the current orchestration process as a template. During the creation of subsequent data processing tasks, you can select a processing template.

      If you select a processing template, the orchestrated processing steps will be deleted.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-69 Selecting a processing template



5. After the processing steps are orchestrated, click **Start Process** to start the processing task.

   After the data processing job is successfully executed, the status changes from **Running** to **Succeeded**, indicating that the data has been processed.

6. After data processing is complete, click **Generate** in the **Operation** column on the **Processing Task** page to generate a processed dataset.

   To view the processed dataset, choose **Data Engineering** > **Data Management** > **Datasets**, and click the **Processed Dataset** tab.

4.5.8 Managing Processing Operators

4.5.8.1 Introduction to Preset Processing Operators

4.5.8.1.1 Text Dataset Processing Operators

The data processing operators provide multiple data operation capabilities, including data extraction, filtering, conversion, and labeling. These operators help you extract useful information from massive data and perform deep processing to generate high-quality training data.

The platform supports processing of text datasets, including data extraction, data conversion, data filtering, and data labeling. **Table 4-29** lists the capabilities of text processing operators.

Table 4-29 Text dataset processing operator capabilities

Category	Operator Name	Operator Description
Data extraction	**Word document content extraction**	Extracts text from a Word document and retain the contents, titles, and body of the original document, but does not retain images, tables, formulas, headers, and footers.
	TXT content extraction	Extracts all text content from a TXT file.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	CSV content extraction	Reads all text content from a CSV file and generates data in JSON format based on the key value of the file content type template.
	PDF content extraction	Extracts text from PDF files and converts the text into structured data. Texts, tables, and formulas can be extracted.
	JSON content extraction	Extracts key-value pair information from a JSON file.
	HTML content extraction	Extracts HTML data content based on the tag path, and deletes other content irrelevant to the tag path to be extracted.
	E-book content extraction	Extracts all text content from the eBook.
Data conversion	**Personal data anonymization**	Anonymizes or directly deletes sensitive personal information, such as mobile numbers, identity documents, email addresses, URLs, license plate numbers in China, IP addresses, MAC addresses, IMEIs, passports, and vehicle identification numbers.
	Symbol standardization	Searches for non-standardized symbols carried in the text for standardization and unified conversion. • Unified space: All Unicode spaces (such as U+00A0 and U+200A) are converted to standard spaces (U+0020). • DBC to SBC: Converts full-width characters in documents to half-width characters. • Punctuation normalization: The following symbols support a unified format: – {"? ": "\?\? "} • Number symbol normalization

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Custom regular expression replacement	Uses the customized regular expression to replace the text content if the data items remain unchanged. The following is an example: • Remove References and the content following References: **\nReferences[\s\S]*** • For the PDF content, remove the content before "0 Introduction". The content before the introduction is irrelevant to knowledge: **[\s\S]{0,10000}0 Introduction** • Delete the content irrelevant to knowledge before "1.1 Introduction to Java" from the PDF file: **[\s\S]{0,10000}1\. 1 Introduction to Java**
	Date and time format conversion	Automatically identifies the date, time, and week, and converts the date, time, and week based on the selected format.
Data filtering	**Filtering abnormal characters**	Searches for abnormal characters carried in each data record in the dataset and replaces the abnormal characters with null values. The data items remain unchanged. • Invisible characters, for example, U+0000-U+001F • Web page tag symbols: <style></style> • Special space: [\u2000-\u2009]
	Custom regular expression filtering	Deletes the data that complies with the customized regular expression.
	User-defined keyword filtering	Deletes data that contains keywords.
	Sensitive word filtering	Automatically detects and filters sensitive data such as pornography, violence, and politics in text.
	Filtering based on the text length	Retains the data within the specified length range based on the configured text length.
	Redundant information filtering	Deletes redundant information from the text based on the paragraph granularity without changing the data items. Examples include figure captions, table captions, and references.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	N-gram feature filtering	Determines the document repetition degree. The repetition of words in a document is calculated based on the N values of features. In this case, the following two algorithms can be used to check whether the result is greater than the feature threshold. If the result is greater than the threshold, the document is deleted. • Top-gram filtering: Calculates the proportion of the **most repeated** grams to the total length. If the proportion is greater than the feature threshold, those grams are deleted. • Gram repetition rate filtering: Calculates the proportion of **all repeated** grams to the total length. If the proportion is greater than the feature threshold, those grams are deleted.
	Paragraph feature filtering	The filtering is based on the following: • Paragraph repetition rate • Proportion of the length of repeated paragraphs • Proportion of non-Chinese characters
	Sentence feature filtering	Uses punctuations in a document as sentence separators and collects statistics on the length of each sentence. If the average length of a document is greater than the configured length, the document is retained. Otherwise, the entire document is deleted. The filtering is based on the following: • Average length of sentences to be retained
	Word feature filtering	Calculates the number of words after a document is segmented based on the system word library. After word segmentation, the total number of words is counted. The average word length is the total length of all words divided by the total number of words. If both the total length and the total number of words are met, the current document is retained. The filtering is based on the following: • Number of words to be reserved • Average length of words to be retained

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Filtering the incomplete sentence at the end of a paragraph	Checks whether the content at the end of a paragraph is complete based on the sentence-level filtering granularity, and filters out the content if the content is incomplete.
	Advertisement data filtering	Deletes a sentence that includes advertisement data from the text, based on a filtering granularity of a sentence.
	Q&A pair filtering	Filters Q&A pairs that meet the following conditions: • The question is not in the string format. • The answer is empty. • The answer is meaningless.
	Language filtering	Obtains the language type of the document based on the language detection model and filters the document in the required language.
	Global Text Deduplication	Detects and removes duplicate or highly similar text from data to prevent model overfitting or reduced generalization.
	General semantics scoring of SFT data	Uses the LLM to check and score the general semantics of SFT data and filter data based on the scoring threshold.
	Rule-based inspection and filtering of SFT data	Checks and filters SFT data based on the selected rule.
	CoT scoring of SFT data	Uses the LLM to check and score the CoT of SFT data and filter data based on the scoring threshold.
Data labeling	**Pre-trained text classification**	Classifies the pre-trained text, such as news, education, and health. The supported languages include Chinese and English.
	General quality assessment	Assesses the general quality of text, such as fluency, clarity, and diversity.
	Problem timeliness evaluation	Checks whether the problem is time-sensitive and provides the reason.
	Answer quality scoring	Scores the quality of answers to fine-tuning datasets, such as logical consistency and fact correctness.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Syntax quality assessment	Assesses the syntax quality of texts, such as relevance and standardization.

Word Document Content Extraction

- Applicable file format: document > docx
- Parameter description:

  Type of content to be extracted: Extracts text from a Word document and retain the contents, titles, and body of the original document, but does not retain images, tables, formulas, headers, and footers.
- Parameter configuration example:

  No parameters need to be set. By default, the contents, title, and body of the original document are retained, and the images, tables, formulas, headers, and footers are not retained.
- Extraction example

  {"fileName":"JAVA from Beginner to Master.docx","text":"JAVA is a cross-platform..."}

TXT Content Extraction

- Applicable file format: document > txt
- Parameter description:

  Type of content to be extracted: By default, the full text is extracted into one line. Text can also be extracted by paragraph. The text is divided into multiple lines based on the entered separator. Each separator is separated by a vertical bar (|). The separator can contain a maximum of 100 characters.
- Extraction example

  {fileName":"TXT file name .txt","text":"This is the first line. "}

CSV Content Extraction

- Applicable file formats: Text > single-turn Q&A, single-turn Q&A (with a persona), and Q&A ranking.
- Parameter description

  Type of content to be extracted: Reads all text content from a CSV file and generates data in JSON format based on the key value of the file content type template.
- Parameter configuration example

  No parameters need to be set.
- Extraction example

  If the extracted CSV content is "Hello, please introduce yourself. I am Pangu model.", the extracted content is {"context":"Hello, please introduce yourself","target":"I am Pangu model."}

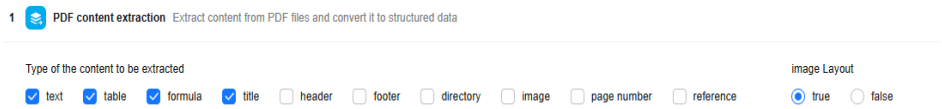
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

PDF Content Extraction

- Applicable file format: document > PDF
- Parameter description

  Type of content to be extracted: By default, the text, tables, formulas, and titles are retained. You can select the type to be saved. The types that are not selected will be removed.

  Refined content extraction: indicates whether to support image content extraction after layout analysis.
- Parameter configuration example



- Extraction example

  {"fileName":"JAVA from Beginner to Master.pdf","text":"JAVA is a cross-platform..."}

JSON Content Extraction

- Applicable file formats: Text > pre-trained text, single turn Q&A, single-turn Q&A (with a system persona), multi-turn Q&A, multi-turn Q&A (with a system persona), Q&A ranking, DPO, and DPO (with a system persona).
- Parameter description

  Type of content to be extracted: Reads the key-value pair information in all text.
- Parameter configuration example

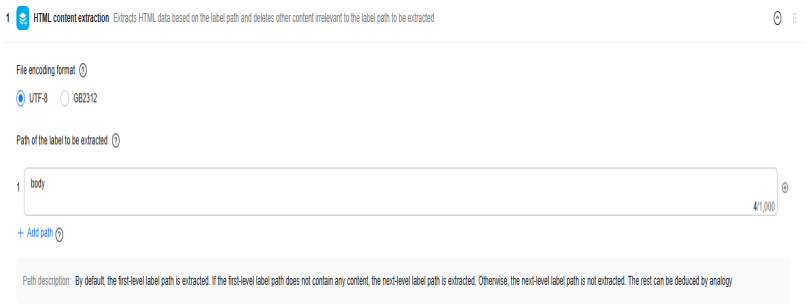
  No parameters need to be set.
- Extraction example

  {"context": "Hello, please introduce yourself.","target":"I am a Pangu model."}

HTML Content Extraction

- Applicable file format: text > web page
- Parameter description

  Type of content to be extracted: The default file encoding format is UTF-8. The GB2312 format is supported. By default, the body is extracted. You can customize the content to be extracted. Multiple labels can be extracted. Labels are separated by commas (,), for example, A, B, C. That is, the content of label A, B, or C is extracted.
- Parameter configuration example



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Extraction example

  {"text":"#\nI am Pangu model.\nPangu model is an advanced AI model that is dedicated to providing intelligent solutions for various industries.\n","fileName":"Web page.html"}

E-book Content Extraction

- Applicable file format: document > mobi/epub
- Parameter description

  Type of content to be extracted: Extracts all text content from MOBI or EPUB e-books.
- Parameter configuration example

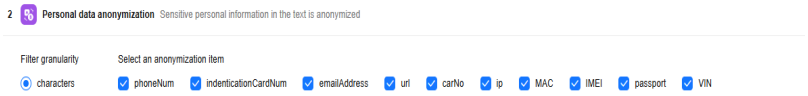
  No parameters need to be set.
- Extraction example

  {"fileName":"JAVA from Beginner to Master.epub","text":"JAVA is a cross-platform..."}

Personal Data Anonymization

- Applicable file format: text
- Parameter description

  Type of content to be converted: Anonymizes sensitive personal information in the text, such as mobile numbers, ID cards, email addresses, URLs, license plate numbers in China, IP addresses, MAC addresses, IMEIs, passports, and vehicle identification numbers. By default, all options are selected. You can also select some of them.
- Parameter configuration example



- Conversion example

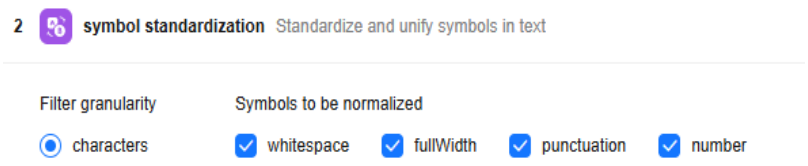
  Before processing: "Data is from www.test.com"

  After processing: "Data comes from*******"

Symbol standardization

- Applicable file format: text
- Parameter description

  Type of content to be converted: Non-standard symbols in the text can be converted to standard symbols. The non-standard symbols include spaces, DBC symbols, punctuations, and number symbols. By default, all non-standard symbols are selected. The filtering granularity is character.
- Parameter configuration example



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Custom Regular Expression Replacement

- Applicable file format: text
- Parameter description

  Type of content to be converted: Uses the customized regular expression to replace the text content if the data items remain unchanged.
- Parameter configuration example



- Conversion example

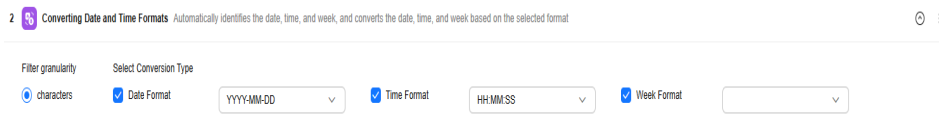
  Before processing: {"text":"This is the main content aeiou in the test aeiou. "}

  After processing: {"text":"This is the main content 11111 in the test 11111. "}

Date and time format conversion

- Applicable file format: text
- Parameter description

  Type of content to be converted: Automatically identifies the date, time, and week, and converts the date, time, and week based on the selected format. The conversion types include date format, time format, and week format. By default, all of them are selected. You can also select some of them.
- Parameter configuration example



- Conversion example

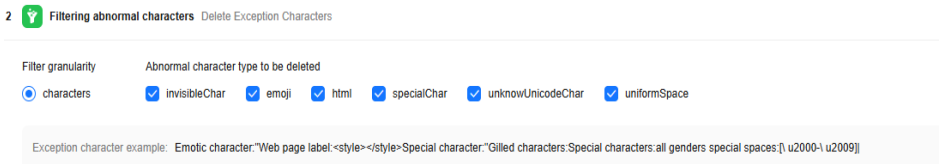
  Before processing: {"text":"Today is Monday, March 3, 2025. The rain is heavy in the morning. "}

  After processing: {"text":"Today is Monday, 2025-03-03 00:00:00. The rain is heavy in the morning. "}

Filtering abnormal characters

- Applicable file format: text
- Parameter description

  Type of content to be converted: Searches for abnormal characters carried in each data record in the dataset and replaces the abnormal characters with null values. The data items remain unchanged. Types of abnormal characters include invisible characters, emojis, web page labels, special characters, garbled characters, and special spaces. By default, all types are selected. You can also select some of them.
- Parameter configuration example



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Filtering example

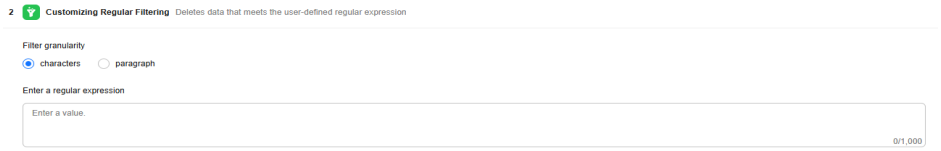
  Before processing: {"text":"Test exception. <style></style>Haha. Limited-time offer! ☺"}

  After processing: {"text":"Test exception. Haha. Limited-time offer!"}

Custom Regular Expression Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Filters content based on a custom regular expression. The filtering granularity can be character (default) or paragraph.
- Parameter configuration example



- Filtering example

  Filtering out the content following "References"

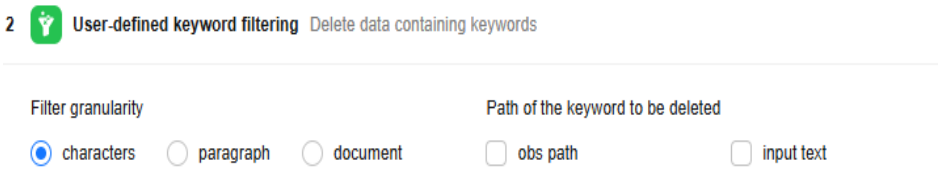
  Before processing: {"text":"This is the body content. References [1] Author 1, Article 1, Journal 1, 2021. [2] Author 2, Article 2, Journal 2, 2022."}

  After processing: {"text":"This is the body content. "}

Custom Keyword Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: The filtering granularity can be character (default), paragraph, or document. The path of the keyword to be deleted supports keyword import from OBS and text input.
- Parameter configuration example



- Filtering example

  Filtering by keyword "test"

  Before processing: {"text":"Keyword test. This is a test data record. "}

  After processing: {"text":"Keyword. This is a test data record. "}

Sensitive Word Filtering

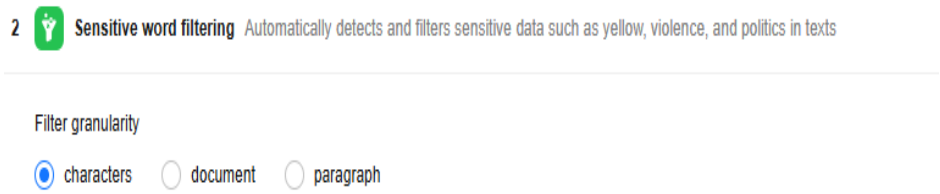
- Applicable file format: text
- Parameter description

  Type of content to be filtered: Automatically detects and filters sensitive data such as pornographic, violent, and political content in the text. Sensitive

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

words need to be preset. The filtering granularity can be character (default), paragraph, or document.

- Parameter configuration example



- Filtering example

  Before processing: {"text":" prostitute test"}

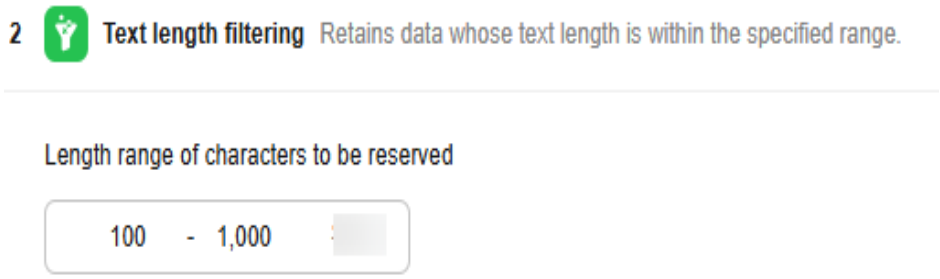
  After processing: {"text":"test"}

Filtering Based on the Text Length

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Retains data within the specified text length. By default, the length of the characters to be reserved ranges from 100 to 1000 characters, which can be modified.

- Parameter configuration example



- Filtering example

  Before processing: {"text":"Test length"}

  After processing: {"text":""}

Redundant Information Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Deletes redundant information from the text based on the paragraph granularity without changing the data items. The content types that can be filtered include figures, notes, and references. By default, all content types are selected. You can also select some of them.

N-gram Feature Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: The filtering granularity is document. You can select top-gram filtering or gram repetition rate filtering. By default, top-gram

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

filtering is selected. In top-gram filtering mode, the default value of feature N is 2, and the default value of feature threshold is 0.18. In gram repetition rate filtering mode, the default value of feature N is 2, and the default value of feature threshold is 0.15. The values can be changed.

- Parameter configuration example



- Filtering example

  Before processing: {"text":"Wake up. Today is Sunday. Today is a holiday. Tomorrow is Monday. Tomorrow is a working day. "}

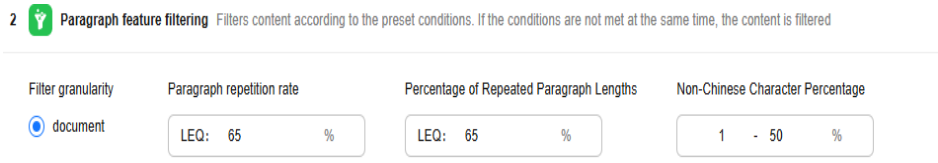
  After processing: {"text":""}

Paragraph Feature Filtering

- Applicable file format: text
- Parameter description

  Types of content to be filtered: Filters the content based on the document filtering granularity, paragraph repetition rate, proportion of the length of repeated paragraphs, and proportion of non-Chinese characters. If any of the specified conditions is not met, the content is filtered out. The default values are as follows: The paragraph repetition rate is less than or equal to 65%, the proportion of the length of repeated paragraphs is less than or equal to 65%, and the proportion of non-Chinese characters ranges from 1% to 50%. The values can be changed.

- Parameter configuration example



- Filtering example

  Before processing: {"text":"It is said that the fox only appears to those with pure hearts and sincere wishes. Under the light of the moon, it will gracefully emerge, gazing at the visitor with eyes that shimmer with wisdom. Only when the fox senses the visitor's sincerity and purity will it speak, asking about their desires. Yet, the granting of a wish is never without cost. Each fulfilled wish demands a corresponding price—perhaps a cherished memory, a beloved possession, or even a piece of one's life. Therefore, villagers must carefully ponder before making a wish, considering whether they are truly willing to pay such a price."} "}

  After processing:

Sentence Feature Filtering

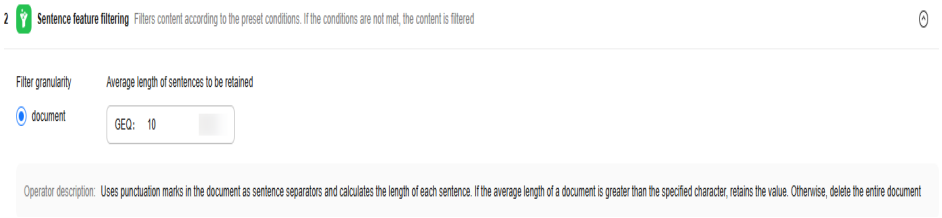
- Applicable file format: text
- Parameter description

  Type of content to be filtered: Filters the content based on the document filtering granularity and the average sentence length to be retained. If the

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

content does not meet the requirements, the content is filtered out. The default average sentence length to be retained is greater than or equal to 10 characters, which can be modified.

- Parameter configuration example



- Filtering example

  Before processing: {"text":"In a small village, there is a legend. According to the legend, a mysterious fox appears in the village forest every full moon night."}

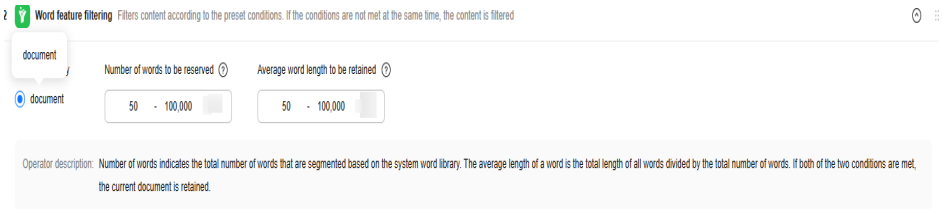
  After processing: {"text":""}

Word Feature Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Filters the content by document based on the number of words to be retained (50-100,000 characters by default) and the average length of words to be retained (50-100,000 characters by default). If any of two conditions is not met, the content is filtered out. The default value can be changed.

- Parameter configuration example



- Filtering example

  Before processing: {"text":"It is said that the fox only appears to those with pure hearts and sincere wishes. " }

  After processing: {"text":""}

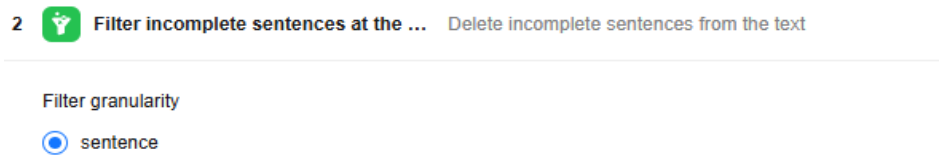
Filtering of the Incomplete Sentence at the End of a Paragraph

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Checks whether the content at the end of a paragraph is complete based on the sentence-level filtering granularity, and filters out the content if the content is incomplete.

- Parameter configuration example

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



- Filtering example

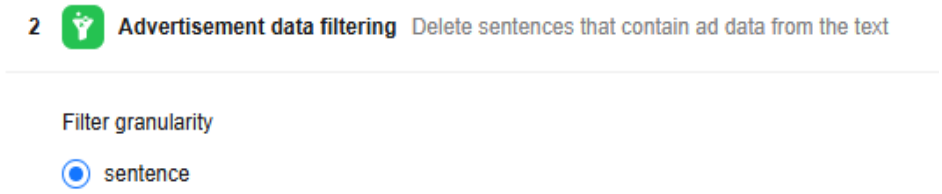
  Before processing: "JAVA is an object-oriented programming language. Use JAVA to,"

  After processing: "JAVA is an object-oriented programming language."

Advertisement Data Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Deletes a sentence that includes advertisement data from the text, based on a filtering granularity of a sentence.
- Parameter configuration example



- Filtering example

  Before processing: {"text":"Specific discount! Buy our products and enjoy a discount of up to 50%! Click the link below to avail the discount at https://example.com. Seize this opportunity now and take action!}

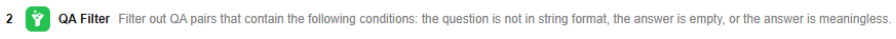
  After processing: {"text":""}

Q&A Pair Filtering

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Filters the Q&A pairs whose questions are not in the string format, answers are empty, or answers are meaningless.
- Parameter configuration example

  No parameters need to be set.



- Filtering example

  Before processing: {"text":"[{"context":"Hello","target":"Yes"), {"context":"list","target":""}]"}

  After processing: {"text":"[{"context":"Hello","target":"Yes"}]"}

Language Filtering

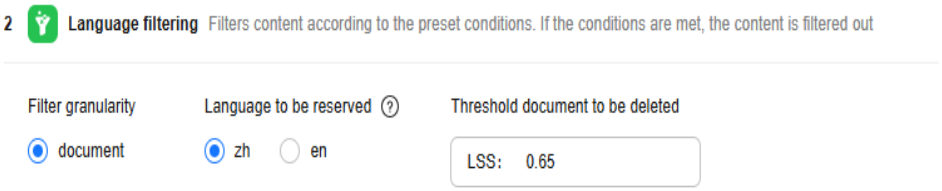
- Applicable file format: text

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Parameter description

  Type of content to be filtered: Filters the content by document based on the language to be retained and the threshold of documents to be deleted. If documents meet filtering criteria, they are filtered out. The language to be retained is Chinese by default. You can select English. The default value of the threshold of documents to be deleted is less than 0.65. The values can be modified.

- Parameter configuration example

- Filtering example

  Before processing: {"text":" Hello, my name is Li Ming. I am excited to introduce myself and share a bit about who l am."}

  After processing: {"text":""}

Global Text Deduplication

- Applicable file format: text
- Parameter description

  Type of content to be filtered: Detects and removes duplicate or highly similar text from data to prevent model overfitting or reduced generalization.

- Parameter configuration example

  No parameters need to be set.

- Filtering example

  Before processing: [{"fileName":"text1.txt","text":"It is said that the fox only appears to those with pure hearts and sincere wishes."}, {"fileName":"text2.txt","text":"It is said that the fox only appears to those with pure hearts and sincere wishes."}

  After processing: {{"fileName":"text1.txt","text":"It is said that the fox only appears to those with pure hearts and sincere wishes."}}

General Semantics Scoring of SFT

- Applicable file formats: Text > single-turn Q&A, single-turn Q&A (with a persona), and Q&A ranking.
- Parameter description

  Type of content to be filtered: Uses the LLM to check and score the general semantics of SFT data and filter data based on the scoring threshold. You need to select a model. The value of the filter field can be **Yes** or **No**. The default value is **No**. The threshold value ranges from 0.0 to 10.0. Samples whose value is less than the threshold will be filtered out. The default threshold is **6**, and the threshold can be changed.

- Filtering example

  Before processing: {"system":"You are an expert in calculations.","context":"Calculate the travel time based on the distance

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

between two cities. Starting city: Beijing, destination city: Shanghai, distance: 1,318 km. \n","target":"To calculate the travel time between two cities based on the distance, we need to consider the mode of travel, as each has a different average speed. Here's a general calculation for common modes of travel between Beijing and Shanghai, which are 1,318 km apart. High-speed train: Approximately 5-6 hours; Plane: Approximately 2 hours; Car: Approximately 13-14 hours. "}{"system":"You are a Q&A expert","context":"Find the number of occurrences of a specific keyword in a given article. \nArticle: Many people will stay up late to watch the matches during this year's World Cup. \nKeyword: World Cup\n","target":"Keyword \"World Cup\" appears once in the article. "}

After processing: {"context":"Find the number of occurrences of a specific keyword in the given article. \nArticle: Many people will stay up late to watch the matches during this year's World Cup. \nKeyword: World Cup \n","filter":0.0,"qa_quality score":{"reason":"The large model correctly identified that the keyword \"World Cup\" appeared once in the article and provided the correct answer. ","score":10.0},"system":"You are a Q&A expert","target":"Keyword \"World Cup\" appears once in the article. "}

Rule-based Inspection and Filtering of SFT Data

- Applicable file formats: Text > single-turn Q&A, single-turn Q&A (with a persona), and Q&A ranking.
- Parameter description

  Type of content to be filtered: Checks the SFT data quality and filters the data based on the selected rule. The filtering rules include character strings, truncated text, incomplete content, mixed Chinese and English characters, mixed Traditional to Simplified Chinese characters, repeated content, special characters, unaligned parentheses, repeated patterns, garbled characters, inconsistent Chinese and English replies, sensitive model identity, and no slow thinking. By default, all filtering rules are selected, you can also select some of them. The filtering field can be set to **Yes** or **No**. The default value is **No**.
- Filtering example

  Before processing: {"context":"Hello, please introduce yourself.","target":"I am Pangu hello world."}{"context:"Which poet is referred to as the Poet Immortal?","target":"Hello! There are Traditional Chinese."}

  After processing:

CoT Scoring of SFT Data

- Applicable file formats: Text > single-turn Q&A, single-turn Q&A (with a persona), and Q&A ranking.
- Parameter description

  Type of content to be filtered: Uses the LLM to check and score the CoT of SFT data and filter data based on the scoring threshold. You need to select a model. The value of the filter field can be **Yes** or **No**. The default value is **No**. The threshold value ranges from 0.0 to 10.0. Samples whose value is less than the threshold will be filtered out. The default threshold is **6**, and the threshold can be changed. The data judgment rule and data scoring rule can be edited. The value can contain a maximum of 1000 characters.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Filtering example

  Before processing: {"context":"Context","targets":["hello","hi","hello"]}

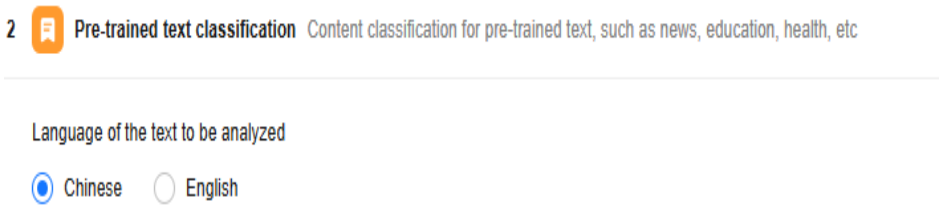
  After processing: {"context":"Context","targets":
  ["hello","hi","hello"],"qa_cot_score":
  {{"result":"Incorrect","score":0.0,"reason"["The model's answer is irrelevant."]},
  {"result":"Incorrect","score":0.0,"reason":"The answer is irrelevant."},
  {"result":"Incorrect","score":0.0,"reason":"[The model's answer is irrelevant.]"}]}

Pre-trained Text Classification

- Applicable file format: text > pre-trained text
- Parameter description

  Type of content to be labeled: Classifies the content of the pre-trained text, for example, news, education, and health. The supported languages include Chinese and English. The default language is Chinese.

- Parameter configuration example

- Example

  {"fileName":"News Labeling Test .docx","text":"Beijing, March 3 (Xinhua) According to the People's Bank of China, the financial market operation in January showed that China issued a total of CNY5,102.75 billion in bonds in January. Of which, CNY1,018.5 billion in government bonds, CNY557.57 billion in local government bonds, CNY704.21 billion in financial bonds, CNY1,279.17 billion in corporate credit bonds, CNY2.73 billion in asset-backed securities, and CNY1,514.78 billion in interbank certificates of deposit. \n As of the end of January, the bond market custody balance in China was CNY178.2 trillion. Of which, the custody balance of the interbank market was CNY156.9 trillion, and that of the exchange market was CNY21.3 trillion. \nAs of the end of January, the custody balance of foreign institutions in the Chinese bond market was CNY4.2 trillion, accounting for 2.3% of the custody balance of the Chinese bond market. Of which, the bond custody balance of foreign institutions in the interbank bond market was CNY4.1 trillion. By bond type, foreign institutions held CNY2.0 trillion of government bonds (48.8%), CNY1.1 trillion of certificates of deposit (25.8%), and CNY0.9 trillion of policy bank bonds (20.8%). \n","pre_classification":"Economy"}

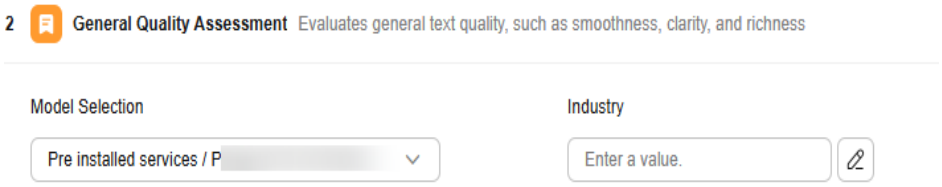
General Quality Assessment

- Applicable file format: text > pre-trained text
- Parameter description

  Type of content to be labeled: Assesses the general quality of text, such as fluency, clarity, and diversity. You need to select a model and an industry. The industry can be manually entered.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Parameter configuration example



- Example

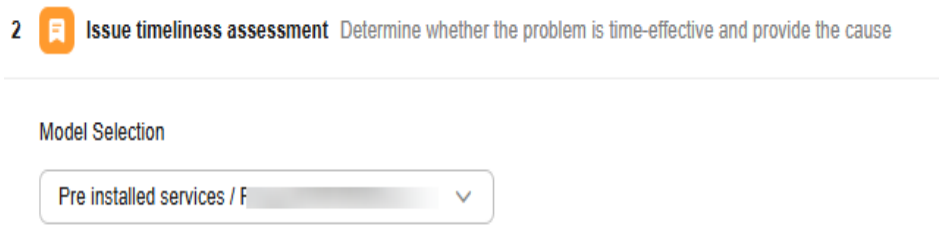
  {"fileName":"News Labeling Test .docx","text":"Beijing, March 3 (Xinhua) According to the People's Bank of China, the financial market operation in January showed that China issued a total of CNY5,102.75 billion in bonds in January. Of which, CNY1,018.5 billion in government bonds, CNY557.57 billion in local government bonds, CNY704.21 billion in financial bonds, CNY1,279.17 billion in corporate credit bonds, CNY2.73 billion in asset-backed securities, and CNY1,514.78 billion in interbank certificates of deposit. \n As of the end of January, the bond market custody balance in China was CNY178.2 trillion. Of which, the custody balance of the interbank market was CNY156.9 trillion, and that of the exchange market was CNY21.3 trillion. \nAs of the end of January, the custody balance of foreign institutions in the Chinese bond market was CNY4.2 trillion, accounting for 2.3% of the custody balance of the Chinese bond market. Of which, the bond custody balance of foreign institutions in the interbank bond market was CNY4.1 trillion. By bond type, foreign institutions held CNY2.0 trillion of government bonds (48.8%), CNY1.1 trillion of certificates of deposit (25.8%), and CNY0.9 trillion of policy bank bonds (20.8%). \n","generalscore":{"Instructiveness":"5","Cleanliness":"5","isIncorrect":"false","Toxicity":"false","Richness":"5","Fluency":"5","knowledge":"5"}}

Problem Timeliness Evaluation

- Applicable file format: text > single-turn Q&A
- Parameter description

  Type of content to be labeled: Determines whether the issue is time-sensitive and provides the reason. You need to select a model.

- Parameter configuration example



- Example

  {"context":"1-1-2 Where is the capital of Jiangsu Province?","target":"Nanjing","timeliness_classification":0}

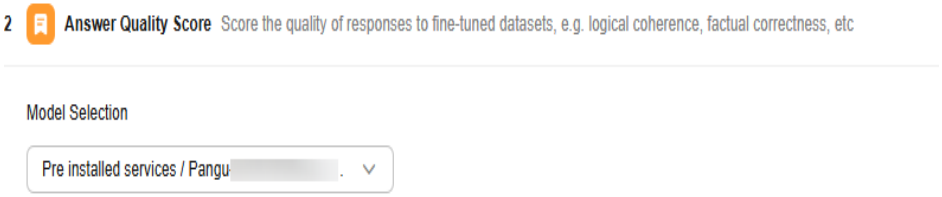
Answer Quality Scoring

- Applicable file format: text > single-turn Q&A
- Parameter description

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Type of content to be labeled: Scores the quality of answers to fine-tuning datasets, such as logical consistency and fact correctness. You need to select a model.

- Parameter configuration example



- Example

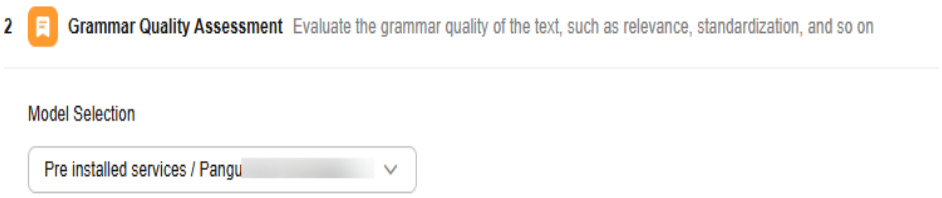
  {"context":"1-2-1 Where is the capital of China? ","target":"Beijing","answer score":{"Logical coherence":10,"Comprehensive score":10,"User requirements satisfied":10,"Completeness":10,"Fact correctness":10}}

Syntax Quality Evaluation

- Applicable file format: text > single-turn Q&A
- Parameter description

  Type of content to be labeled: evaluates the syntax quality of the text, for example, relevance and standardization. You need to select a model.

- Parameter configuration example



- Example

  {"target":"Beijing","context":"1-2-1 Where is the capital of China?","grammar score":{"Unrelated reply":0,"Fact error":0,"Logical error":0,"Non-standard language":0,"Sentence truncation":0,"Improper multi-language mixing":0,"Meaningless repetition":0}}

4.5.8.1.2 Video Dataset Processing Operators

The data processing operators provide multiple data operation capabilities, including data extraction, filtering, conversion, and labeling. These operators help you extract useful information from massive data and perform deep processing to generate high-quality training data.

The platform supports processing of video datasets, including data extraction, data filtering, and data labeling. **Table 4-30** lists the capabilities of video dataset processing operators.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-30 Video dataset processing operator capabilities

Category	Operator Name	Operator Description
Data extraction	**Scene splitting**	Splits a long video into short video clips based on the scene change. If the length of a clip exceeds the specified time threshold, the clip is further split by duration.
Data conversion	**Adding Watermarks**	Adds a full-screen text watermark to a video.
	Video Cropping	Video cropping is to crop unnecessary elements in a video, such as subtitles, logos, watermarks, borders, and dense text, and filter out video files whose area ratio after cropping exceeds the preset threshold. Before using this function, you need to execute the subtitle, logo, watermark, border, and dense text recognition operators.
Data filtering	**Video Metadata Filtering**	Filters videos based on the video metadata (frame rate, resolution, and video duration) and retains only the videos that meet the specified conditions. Note: The standard frame rate of a movie is 24 FPS or 30 FPS.
	Video Aspect Ratio Filtering	Filters videos based on the aspect ratio. The aspect ratio is a ratio of a width to a height of a video image.
Data labeling	**Pornographic Video Detection**	Labels pornographic content.
	Violent and Terrorism Video Detection	Labels violent and terrorism content.
	Political Video Content Detection and Scoring	Labels political content.
	Motion Range Scoring	Calculates and scores the motion range of each pixel in each frame, and identifies videos with too fast motion (for example, > 100 optical flows) or too slow motion (for example, ≤ 2 optical flows). A larger value indicates faster motion.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Quality score	Scores the basic video quality (definition, brightness, blur, image shaking and ghosting, overexposure in low light, and glitch). The value range is (0, 1). A higher value indicates better quality. A video whose score is greater than 0.05 is considered as a video with high basic quality.
	Aesthetics Scoring	Scores the aesthetics of a video from the following dimensions: content (attractive and clear), composition (good object position), color (vital and pleasant), light (obvious contrast), and track (continuous and stable). The value range is (0, 1). A higher value indicates better aesthetics. A video whose score is greater than 0.95 is considered as a video with high aesthetics.
	Watermark Identification	Identifies whether a video contains watermarks.
	Subtitle Identification	Identifies whether a video contains subtitles.
	Logo Identification	Identifies whether a video contains a logo.
	Black Bar Identification	Checks whether a video contains black bars.
	Dense Text Identification	Identifies whether a video contains dense text. A video in which the proportion of dense text area exceeds the specified proportion is a video with dense text. By default, a video with a cropping area proportion greater than or equal to 7% is a video with dense text.
	Video classification (InterVideo2)	Returns video label categories through the operator. There are seven categories at L1 and 700 categories at L4.
	Video Synopsis Generation (Simplified)	Extracts frames from a video and generates a simplified video synopsis through model inference.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Video Synopsis Generation (Detailed)	Extracts frames from a video and generates a detailed video synopsis through model inference.

Scene Splitting

- Applicable file format: video > mp4/avi.
- Parameter description:

  Video to be split: Videos that meet the resolution, duration, and frame rate criteria are split.

  Specifications after video splitting: The maximum duration of a single video slice can be customized. If the duration of the first split slice exceeds the specified value, the video will be further split. The final split result is less than or equal to the specified threshold.

Adding Watermarks

- Applicable file format: video > mp4/avi.
- Parameter description:

  Watermark text: string-type text to be added to the full screen

Video Cropping

- Applicable file format: video > mp4/avi.
- Parameter description:

  Items to be cropped: Remove useless information such as subtitles, logos, watermarks, borders, and dense text from videos.

  Cropping area ratio: The ratio of the cropped video area to the original video area is the cropping area ratio. Videos whose cropping area ratio exceeds the preset threshold will be filtered out. The default value is 30%.

Video Metadata Filtering

- Applicable file format: video > mp4/avi.

Video Aspect Ratio Filtering

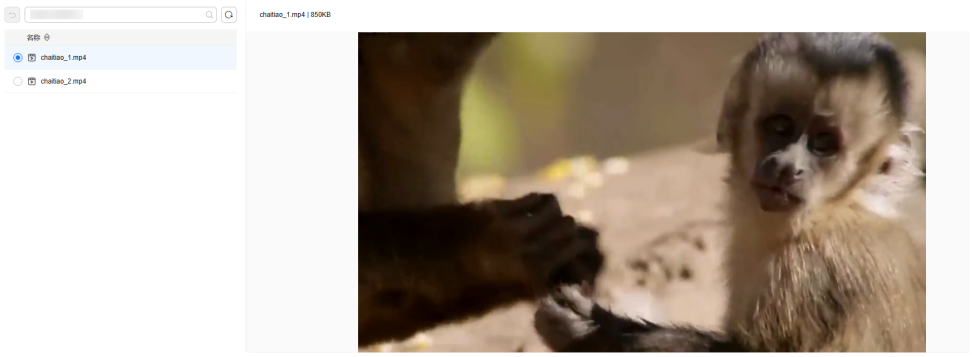
- Applicable file format: video > mp4/avi.
- Parameter description:

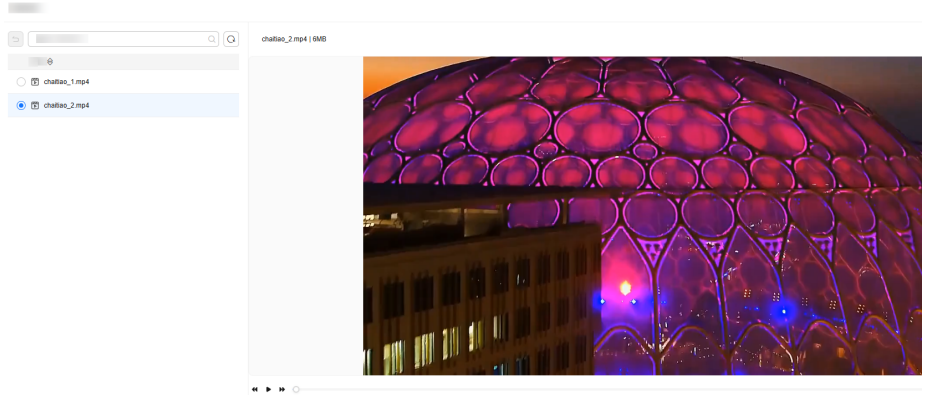
  Aspect ratio threshold: Videos whose aspect ratio exceeds the threshold will be filtered out. The threshold range is (1, 10). You can enter one decimal place.
- Example:

  Original video dataset:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

There are two videos, and their respective aspect ratios are 1.77 and 1.79.





Set the aspect ratio threshold to 1.78. After operator processing, the result is as follows.



Only the video whose aspect ratio is 1.79 is retained.

Pornographic Video Detection

- Applicable file format: video > mp4/avi.
- Operator description: Labels pornographic content.
- Parameter configuration example:

  No parameters need to be set.
- Detection example:

  The results are stored in the annotation file as the video_anti_porn object.

  suggestion: indicates whether the file passes the check. **pass** indicates that the file passes the check and no problem occurs. **review** indicates that

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

manual review is required. You can choose to bypass or block the file based on your review policy. **block** indicates that the file to be reviewed is problematic.

confidence: detection confidence of the model. (Note that the confidence indicates the confidence of the model-provided suggestions.) If **suggestion** is **pass**, the value is 0. If **suggestion** is **review** or **block**, the value ranges from 0 to 1.

label: label of the pornographic content detected by the model. If no pornographic content is detected, the value is empty.



Violent and Terrorism Video Detection

- Applicable file format: video > mp4/avi.
- Operator description: Labels violent and terrorism content.
- Parameter configuration example:

  No parameters need to be set.
- Detection example: The results are stored in the annotation file as the video_anti_terrorism object.

  suggestion: indicates whether the file passes the check. **pass** indicates that the file passes the check and no problem occurs. **review** indicates that manual review is required. You can choose to bypass or block the file based on your review policy. **block** indicates that the file to be reviewed is problematic.

  confidence: detection confidence of the model. (Note that the confidence indicates the confidence of the model-provided suggestions.) If **suggestion** is **pass**, the value is 0. If **suggestion** is **review** or **block**, the value ranges from 0 to 1.

  label: label of the violent and terrorism content detected by the model. If no violent or terrorism content is detected, the value is empty.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



```
{
  "file_name": "terrorism_bad.mp4",
  "video_anti_terrorism": {
    "results": [
      {
        "category": "terrorism",
        "confidence": 0.9961,
        "label": "bloody"
      }
    ],
    "suggestion": "block"
  },
  "time": "00:00:16"
}
```

Political Video Content Detection and Scoring

- Applicable file format: video > mp4/avi.
- Operator description:

  Labels political content.
- Parameter configuration example:

  No parameters need to be set.
- Detection example:

  The results are stored in the annotation file as the video_anti_politics object.

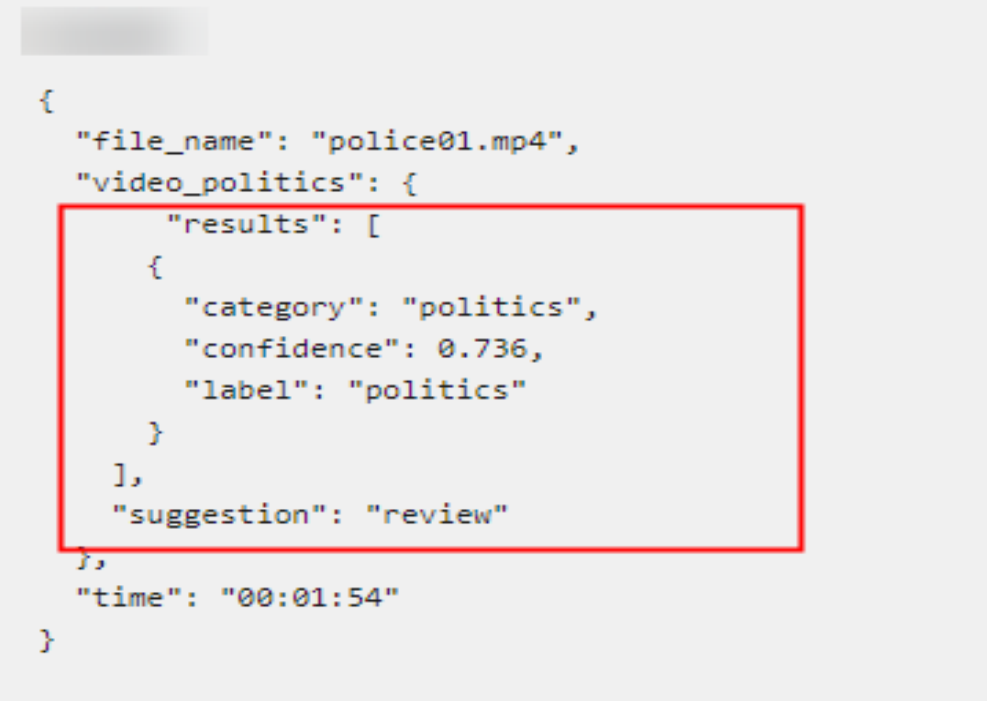
  suggestion: indicates whether the file passes the check. **pass** indicates that the file passes the check and no problem occurs. **review** indicates that manual review is required. You can choose to bypass or block the file based on your review policy. **block** indicates that the file to be reviewed is problematic.

  result: result returned by the model after file detection, including the suggestion, confidence, and label. One or more records can be returned.

  confidence: detection confidence of the model. (Note that the confidence indicates the confidence of the model-provided suggestions.) If **suggestion** is **pass**, the value is 0. If **suggestion** is **review** or **block**, the value ranges from 0 to 1.

  label: label of the political content detected by the model. If no political content is detected, the value is empty.

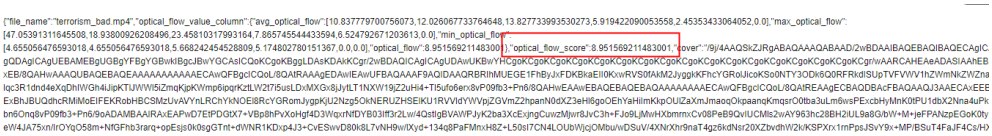
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



Motion Range Scoring

- Applicable file format: video > mp4/avi.
- Scoring description:

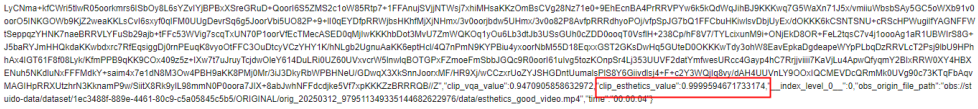
  Identifies videos with too fast or too slow motion. A larger value indicates faster motion. If the motion range is greater than 100 optical flows, the motion is too fast. If the motion range is less than or equal to 2 optical flows, the motion is too slow.
- Scoring example: The motion range score is displayed in the JSONL file.


{"file_name":"terrorism_bad.mp4","optical_flow_value_column":{"avg_optical_flow":[10.837779700756073,12.026067733764648,13.827733993530273,5.919422090053558,2.45353433064052,0.0],"max_optical_flow":
[47.05391311645508,18.93800926208496,23.45810317993164,7.865745544433594,6.524792671203613,0.0],"min_optical_flow":
[4.655056476593018,4.655056476593018,5.668242454528809,5.174802780151367,0.0,0.0],"optical_flow":8.951569211483001},"optical_flow_score":8.951569211483001,"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgIC
gQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLDAsKDAkKCgr/2wBDAQICAgICAgUDAwUKBwYHCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCAHEAeADASIAAhEB
xEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQAAAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZna
lqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6erx8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAQAAQJ3AAECAxEEE
ExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk
bn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwD7EtPDDYBX03lft3r2Lw/4QstgBVAVPJyK2ba3XcExjngCuwzMjwr8JvC3h/wAZ6v4lt

Aesthetics Scoring

- Applicable file format: video > mp4/avi.
- Scoring description:

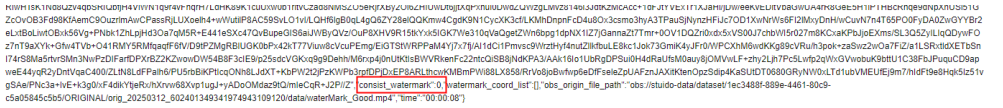
  Scores the aesthetics of a video from the following dimensions: content (attractive and clear), composition (good object position), color (vital and pleasant), light (obvious contrast), and track (continuous and stable). The value range is (0, 1). A higher value indicates better aesthetics. A video whose score is greater than 0.95 is considered as a video with high aesthetics.
- Scoring example: The aesthetics scores are stored in a JSONL file as the clip_esthetics_value object.


LyCNma+kfCWri5ttwR05oorkmrs6iSbOy8L6sY2vIYjBPBtxXSreGRuD+QoorI6S5ZMS2c1oW85Rtp7+1FFAnujSVjNTWsj7xhiMHsaKKzOmBsCVg28Nz71e0+9EhEcnBA4PtRRVPYw6k5kQdWqJihBJ9KKKwq7G5lWaXn71J5xvmiuWbsbSAy5GC5oWXb91v0
oorO5iNKGOWb9KjZ2weaKKLsCvl6sxyf0qIFM0UUgDevrSq6g5JoorVbl5UO82P+9+iI0qEYDfpRRWjbsHKhfMjXjNHmx/3v0oorjbdv5UHmx/3v0o82P8AvfpRRRdhyoPOjvfpSpJG7bQ1FFCbuHKlwlsvDbjUyEx/dOKKK6kCSNTSNU+cRScHPWuglfYAGNFFW
tSeppqzYHNK7naeBRRVLYFuSb29ajb+tFFc53WVlg7scqTxUN70P1oorVfEcTMecASED0qMjlwKKKhbDot3MvU7ZmWQKOq1yOu6Lb3dtJb3USsGUh0cZDD0ooqT0VsfiH+238C/hF8V7/TYLcixunM9i+ONjEkD8OR+FeL2tqsC7v4j1oooAg1aR1UBWIrS8G+
J5baRYJmHHQkdaKKwbdxrc7RfEgsiggDj0rnPEuqK8vyoOtFFC3OuDtcyVCzYHY1K/hNLgb2JgnuAaKK6eptHcl/4Q7nPmN9KYPBiu4yxoorNbM55D18EqxxGST2GKsDwHq5GUteD0OKKKwTdy3ohW8EavEpkaDgdeapeWYpPLbqDzRRVLcT2Psj9lbU9HPh
hAx4IGT61F8f08Lyk/KfmPPB9qKK9COx409z5z+IXw7t7uJruyTcjdwOleY614DuLRi0UZ60UVxvcrW5lnwlqBOTGPxFZmoeFmSbbJGQc9R0oorI61ulvg5tozKOnpSr4Lj353UUVF2datYmfvesURcc4Gayp4hC7RrjjviiI7KaVjLu4ApwQfyqmY2BlxRRW0XY4HBX
ENuh5NKdluNxFFFMdkY+saim4x7e1dN8M3Ow4PBH9aKK8PMj0Mr/3iJ3DkyRbWPBHNeU/GDwqX3XkSnnJoorxMF/HR9Xj/wCCzxrUoZYJSHGDmtUumalsPtS8Y8GiivdisI4+F+c2Y3WQjlq8vy/dAH4UUVnLY9OOxIQCMEVDcQRmMk0UVg90c73KTqFbAqv
MAGIHpRRXUtzhrN3KknamP9w/SitX8Rk9yIL98mmN0P0oora7JIX+8abJwhNFFdcdjke5Vf7xpKKKZzBRRRQB//Z","clip_vqa_value":0.9470905858632972,"clip_esthetics_value":0.9999594671733174,"__index_level_0__":0,"obs_origin_file_path":"obs://stuido-data/dataset/1ec3488f-889e-4461-80c9-c5a05845c5b5/ORIGINAL/orig_20250312_979511349335144682622976/data/esthetics_good_video.mp4","time":"00:00:04"}

Watermark Identification

- Applicable file format: video > mp4/avi.
- Operator description:

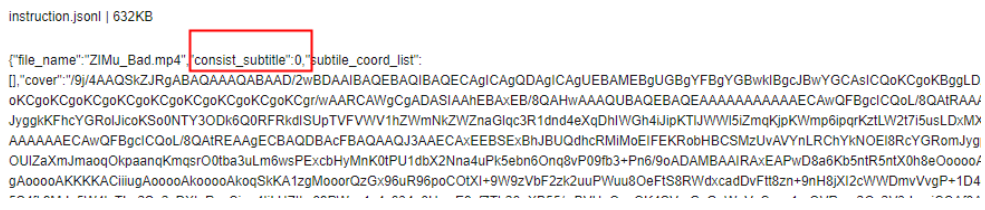
  Identifies whether a video contains watermarks.
- Example: The JSONL file shows whether the watermark is identified. If the value of **consist_watermark** is 1, the watermark is identified. If the value is 0, no watermark is identified.


ZcOvOB3Fd98KfAemC9OuzrlmAwCPassRjLUXoelh4+wWutiIP8AC59SvLO1vl/LQHf6lgB0qL4gQ6ZY28elQQKmw4CgdK9N1CycXK3cf/LKMhDnpnFcD4u8Ox3csmo3hyA3TPauSjNynzHFiJc7OD1XwNrWs6FIZlMxyDnH/wCuvN7n4T65PO0FyDA0ZwGYYBr2
eLxtBoLiwtOBxk56Vg+PNbk1ZnLpjHd3Oa7qM5R+E441eSXc47QvBupeGlS6aiJWByQVz/OuP8XHV9R15tkYxk5IGK7We310qVaQgetZWn6bpg1dpNX1lZ7jGannaZt7Tmr+0OV1DQZri0xdx5xVS00J7chbWI5r027m8KCxaKPbJjoEXms/SL3Q5ZyILlqQDywFO
z7nT9aXYk+Gfw4TVb+O41RMY5RMfqaqf68fV/D9tPZMgRBIUGK0bPx42kT77Viuw8cVcuPEmg/EiGTStWRPPaM4Yj7x7fj/A1fdCi1Pmvsc9WrztHyf4nutZIlkfbuLE8kc1Jok73GmiK4yJFr0WPCXhM6wdKKg89cVRu/h3pok+zaSwz2wOa7FiZ/a1LSRxtldXETbSn
I74rSdMa5rtvrSMn3NwPzDIFarfDPXrBZ2KZwowDW54B8F3cIE9/p25sdcVGKxq9g9Dehh/M6rxp4j0nUtKIlsBWVRkenFc22ntcQiSB8jNdKPA3/AAk16lo1UbRgDPSui0H4dRaUfsM0auy8jOMVwLF+zhy2Ljh7Pc5Lwfp2qWxGVwobuK9bttU1C38FbJPuquCD9ap
weE44yqR2yDntVqaC400/ZLtN8LdFPalh6/PU5rbBiKPticqONh8LJdXT+KbPW2t2jPzKWPb3rpfDPjDxEP8ARLthcwKMBmPWi88LX858/RrVo8joBwfwp6eDfFseleZpUAFznJAXitKtenOpzSdip4KaSUtDT0680GRyNW0xLTd1ubVMEUfEj9m7/hIdFt9e8Hqk5lz51v
gSAe/PNc3a+IvE+k3g0/xF4dikYtjeRx/hXrvw68Xvp1ugJ+yADoOMdaz9tQ/mleCqR+J2P//Z","consist_watermark":0,"watermark_coord_list":[],"obs_origin_file_path":"obs://stuido-data/dataset/1ec3488f-889e-4461-80c9-
c5a05845c5b5/ORIGINAL/orig_20250312_602401349341974943109120/data/waterMark_Good.mp4","time":"00:00:08"}

Subtitle Identification

- Applicable file format: video > mp4/avi.
- Operator description:

  Identifies whether a video contains subtitles.
- Example: The JSONL file shows whether the subtitle is identified. If the value of **consist_subtitle** is 1, the subtitle is identified. If the value is 0, no subtitle is identified.


instruction.jsonl | 632KB

{"file_name":"ZIMu_Bad.mp4","consist_subtitle":0,"subtitle_coord_list":
[],"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgICAgQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLD
oKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCAWgCgADASIAAhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAA/
JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWWl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMX
AAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAQAAQJ3AAECAxEEBSExBhBJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJyg
OUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwD8a6Kb5ntR5ntX0h8eOoooo/
gAooooAKKKKACiiiugAooooAkooooAkoqSkKA1zgMooorQzGx96uRw96poCOtXI+9W9zVbF2zk2uuPWuu8OeFtS8RWdxcadDvFtt8zn+9nH8jXI2cWWDmvVvgP+1D4

Logo Identification

- Applicable file format: video > mp4/avi.
- Operator description:

  Identifies whether a video contains a logo.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

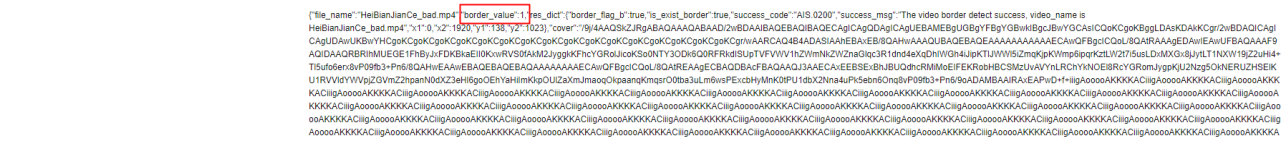
- Example: The JSONL file shows whether the logo is identified. If the value of **consist_logo** is 1, the logo is identified. If the value is 0, no logo is identified.


NSD4reL7i5zqUwA/8aGiiun+zcvl/CqfXcR3Hr8d/iEEY2HWZuv980p+PXiAjA1mb/AL7NFFcf9i5ff+EiPrNZ9TmNRTdbx2lickscncapNNI3G4/hRRX0VM5029SLe/8AeP50vnP7UUUXYDsn1NLvPoKKKwbdzMOyEdTXY/Dn4u994AkT+zbhowr7wVYqQfYjpRRXm
YvD0sTDkqK6NYjEj1r/AIbd8qjSSJdXucA4P+kmnaF+274h8wF9IVucjk4maiivI5cK5G4v90fVQWiOyt/289fa3wL66x/18N/hVfeft4+lt+vvLg8H7Q3+FFFeRDhHIXJ/uvxPQUmaeh/8FCvGquFGpz8dMzGuil/4KFeNUUMNUmwP+mhoorDFcJ5Emv3QOckrkh/4KO
eNiTaNVn/7/Gq4/wCCyjNJMnUJTg9TlaKKwp8I5Cn/C/EwWLrd/wL4/4KceN1i2nUrj/v6ary/wDBTnxwoLCpcD/ALaNRRW0eEchb/hfiweLrpb/AIlsaH/wVY8c2MvmNqdwQpzjzDXaeH/+C13jK0XyZLaZwowQH60UVo+D8hTdqW3mzyfrmlb3NeP/AILbaA0oTxJ8P
FdT1bZnNaq/8Fc2UvEsH/FS/DaHcRyPs/Ofyoor1lcPYHlunJ9vG9+ZXZ4T8fw8rD9mb4i6t9o8H+B5LLAzkYyn09TWI4Q/bV/YjtNC8jVvgLeG9A/4+thIP6UUVUcgwTi/el/4EAvhb/gpn8KPhz4l2+F/g9YnTb4gDMYz7dRzXrhX/gs14G8OoAnwqsUUjGVgXj/x2iiuf
FcP4G0dZf8AgTEdPF/wXE8Aysl4/CblX6AQLgtpWhH/AMFldFnQSR+HyAwAGwUUV4uMyHBwtaU//Ah7j2/4LC2CJ5sejA46Ajltvwb/AMFhPCeqMY9b8J2LAD7zLz/KiivJlI0IwcIUlp5iXxHWjjqgn8FZV3Hw5ZZPbyR/hUbf8FOPgxNyPDdnk/8ATAf4UUV59erjlpHE
TXo/+AdXMw7/AlKG/DCZvjwza4P/AEyWtCH9vb4XMqvJ4ZswD1zCvFFFePLG5pGX+81P/Av+APmNhP27/wBnQc4Og2BDdzarz+IFFdFTITN6C9yp+AuazHy/UfSySQGR/B2mkHqPsa8/pXkPxc/4KCfskeBZzoT/DHTmwx3YfVG
7p/s0UV28OZvj8hj+WrLTUlyF6Hz9+0Z+3p+xD8RPh5c+G5fgJY3V45zbTwlY3tzjqD5fj8qiP/tVtS0qK/kTTHcw7iY/M+9tzwD+FFFfsmCppJO583iNyIsJFDgcEUw5bgc470UV9VSV0cctmJgeidLp//AB7L9KKK0Cn1J8A9RSYHoKKKDQMD0FGxfSiiteREczE2D1
NG0DmiijkRMncY0zqdoXOPek8+T+5+tFFS4ogfWBJuFJ9a7PRzthXbxwaKKSOg0mlkx97v6VyvjB2LYJ7iiihbAc4eTmiipAjooooAqfrCKO+W2k4BHWt/WPB3iXR7ZLvU/Dt9bQyHEc9xaOiN9CRzRRWF3cDIKkHBFNFJleQf0oorcD/2Q==
_coord_list":[[73,75,291,115],[98,56,287,117]],"obs_origin_file_path":"obs://stuido-data/dataset/c87c00b7-e857-4a8f-aac4-19acd002c808/ORIGINAL/[redacted]orig_20250310_723381348701144670670848/data/waterMark_bad1.mp4","time":"00:00:08"}

Black Bar Identification

- Applicable file format: video > mp4/avi.
- Operator description:

  Identifies whether a video contains black bars.
- Example: If **border_value** is 1, black bars are identified. If **border_value** is 0, black bars are not identified.


{"file_name":"HeiBianJianCe_bad.mp4","border_value":1,"res_dict":{"border_flag_b":true,"is_exist_border":true,"success_code":"AIS.0200","success_msg":"The video border detect success, video_name is
HeiBianJianCe_bad.mp4","x1":0,"x2":1920,"y1":138,"y2":1023},"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgICAgQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLDAsKDAkKCgr/2wBDAQICAgICAgUDAwUDCgcGBwoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCAQ4B4ADASIAAhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEDAwIEAwUFBAQAAAF9
AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWXl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMXGx8jJytLT1NXW19jZ2uHi4+Tl5ufo6er/8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAAEAAF3AAECAxEEBSExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwDwLVNifuE+6MfKPSn/Zf+mQ/75q
ZY822AA52CIWADjh2r90Pysg+y/8ATIf980fZf+mQ/wC+as+VH/d/Wjyjo/wC7+tAFb7L/0yH/AHzR9l6ZD/vmrPiR/3f1o8qP+7+tAFb7TOP8AnPyo+zj/niPyqfA9BRIipXQBT8pO8aj8KTyh/zvX8qZr6UbF9KDMi8of8AOqTxJ/6yvlNfHZX/pKP+as+VH/AHf1o8qP+7+tAFX7OP8AnPyo+zj/niPyqfA9BRIipXQBT8pO8aj8KTyh/zvX8qZr6UbF9KDMi8of8A
PFfypTEnaNT+FStkF9KNi+IAERhjPiLiDwAEb8mMfdiT8FqUgHqKkIB6ik3YCLyEwCeCiR5Cf88F/Kpd+lGxfSoAi8hP+eCi/UfD/nlv5vbyv939ajyv939aiqIDfsq/88F/75o+yrzxvXvmrO0f3f0o2fTP8UXJuV/sqfBAPJF+aPsqfBAPJF+an8sf5TR5Yumsst9ti9ni54L/
3zSfZolvlDnkh/FPorpNBvkJzvrTv2KzkEQf9Smf90VMkft98TvLP90Vjcmmtfh/54v/98mxvrQ8uuY++RU0kZ4+UU3yzdFVASsN8mH/min/fAp3tt+6KPL990Uc7hZdyHyIdvbyY+f9kUeRD/AM8YlwDvkVN5ZuijyzI/AHRV6FaF+SGHj9ym/faqSOGHn9y
nfAp7LmnKuaip0FIT7LDlwA8k/74FH2SH/nlH/3yKk8s/wB0UeWf7oMlgMEJ/5Yp/3wKBCP+WKf98Cp5E6f40Rp1/xoAn8mH/min/fAqz9kb0/Wj7l392gCD7LD/wA8k/74FH2SH/nlH/3yKs+Q39jyG/uCgCp5EP/ADxj/wC+RS+kV/8APJP++RS+WI7o

Dense Text Identification

- Applicable file format: video > mp4/avi.
- Parameter description:

  Proportion of dense text area: A video in which the proportion of dense text area exceeds the specified proportion is a video with dense text. By default, a video with a cropping area proportion greater than or equal to 7% is a video with dense text.
- Example: In the JSONL file, if the value of **consist_densetext** is 1, dense text is identified. If the value is 0, dense text is not identified.


{"dense_text_coord_list":
[],"file_name":"MiJiWenZi_good.mp4","densetext_rate_colum":0.0,"consist_densetext":0,"cover":"/9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAIBAQEBAQIBAQECAgICAgQDAgICAgUEBAMEBgUGBgYFBgYGBwkIBgcJBwYGCAsICQoKCgoKBggLDAsK
DAkKCgr/2wBDAQICAgICAgUDAwUKBwYHCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgoKCgr/wAARCALQBQADASIAAhEBAxEB/8QAHwAAAQUBAQEBAQEAAAAAAAAAAAECAwQFBgcICQoL/8QAtRAAAgEC
DAwIEAwUFBAQAAAF9AQIDAAQRBRIhMUEGE1FhByJxFDKBkaEII0KxwRVS0fAkM2JyggkKFhcYGRolJicoKSo0NTY3ODk6Q0RFRkdISUpTVFVWV1hZWmNkZWZnaGlqc3R1dnd4eXqDhIWGh4iJipKTlJWXl5iZmqKjpKWmp6ipqrKztLW2t7i5usLDxMXGx
8jJytLT1NXW19jZ2uHi4+Tl5ufo6er/8vP09fb3+Pn6/8QAHwEAAwEBAQEBAQEBAQAAAAAAAAECAwQFBgcICQoL/8QAtREAAgECBAQDBAcFBAAEAAF3AAECAxEEBSExBhJBUQdhcRMiMoEIFEKRobHBCSMzUvAVYnLRChYkNOEl8RcYGRomJygpKjU2Nzg5OkNERUZHSElKU1RVVldYWVpjZGVmZ2hpanN0dXZ3eHl6goOEhYaHiImKkpOUlZaXmJmaoqOkpaanqKmqsrO0tba3uLm6wsPExcbHyMnK0tPU1dbX2Nna4uPk5ebn6Onq8vP09fb3+Pn6/9oADAMBAAIRAxEAPwDwLVNifuE+6MfKPSn/Zf+mQ/75q
ZY822AA52CIWADjh2r90Pysg+y/8ATIf980fZf+mQ/wC+as+VH/d/Wjyjo/wC7+tAFb7L/0yH/AHzR9l6ZD/vmrPiR/3f1o8qP+7+tAFb7TOP8AnPyo+zj/niPyqfA9BRIipXQBT8pO8aj8KTyh/zvX8qZr6UbF9KDMi8of8AOqTxJ/6yvlNfHZX/pKP+as+VH/AHf1o8qP+7+tAFX7OP8AnPyo+zj/niPyqfA9BRIipXQBT8pO8aj8KTyh/zvX8qZr6UbF9KDMi8of8A
PFfypTEnaNT+FStkF9KNi+IAERhjPiLiDwAEb8mMfdiT8FqUgHqKkIB6ik3YCLyEwCeCiR5Cf88F/Kpd+lGxfSoAi8hP+eCi/UfD/nlv5vbyv939ajyv939aiqIDfsq/88F/75o+yrzxvXvmrO0f3f0o2fTP8UXJuV/sqfBAPJF+aPsqfBAPJF+an8sf5TR5Yumsst9ti9ni54L/
3zSfZolvlDnkh/FPorpNBvkJzvrTv2KzkEQf9Smf90VMkft98TvLP90Vjcmmtfh/54v/98mxvrQ8uuY++RU0kZ4+UU3yzdFVASsN8mH/min/fAp3tt+6KPL990Uc7hZdyHyIdvbyY+f9kUeRD/AM8YlwDvkVN5ZuijyzI/AHRV6FaF+SGHj9ym/faqSOGHn9y
nfAp7LmnKuaip0FIT7LDlwA8k/74FH2SH/nlH/3yKk8s/wB0UeWf7oMlgMEJ/5Yp/3wKBCP+WKf98Cp5E6f40Rp1/xoAn8mH/min/fAqz9kb0/Wj7l392gCD7LD/wA8k/74FH2SH/nlH/3yKs+Q39jyG/uCgCp5EP/ADxj/wC+RS+kV/8APJP++RS+WI7o

Video Classification (InterVideo2)

- Applicable file format: video > mp4/avi.
- Operator description:

  Automatically classifies short video content and generates corresponding tags.
- Parameter configuration example:

  No parameter configuration required
- Example of category labeling:

  The following information is displayed in the description:

  category_L1_cn: first-level category

  category_L4_cn: fourth-level category

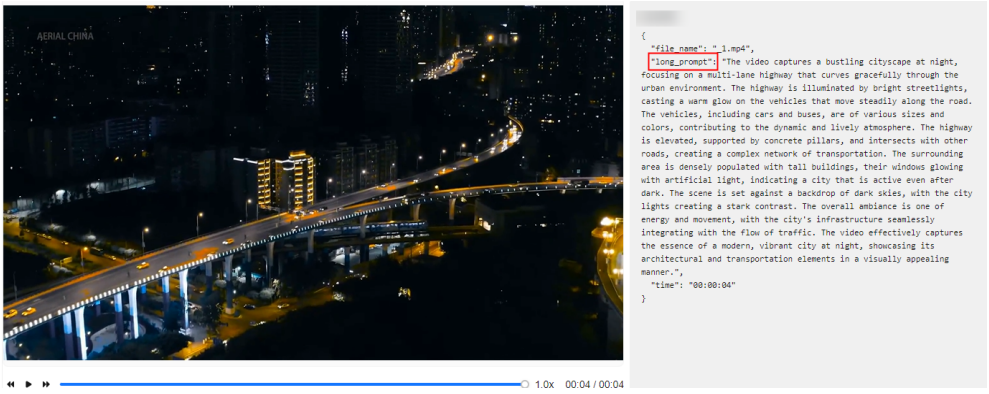
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Video Synopsis Generation (Detailed)

- Applicable file format: video > mp4/avi.
- Operator description:

  Extracts frames from a video and generates a detailed video synopsis through model inference.
- Parameter configuration example:

  No parameter configuration is required.
- Example: The **long_prompt** field in the description indicates the detailed video synopsis.



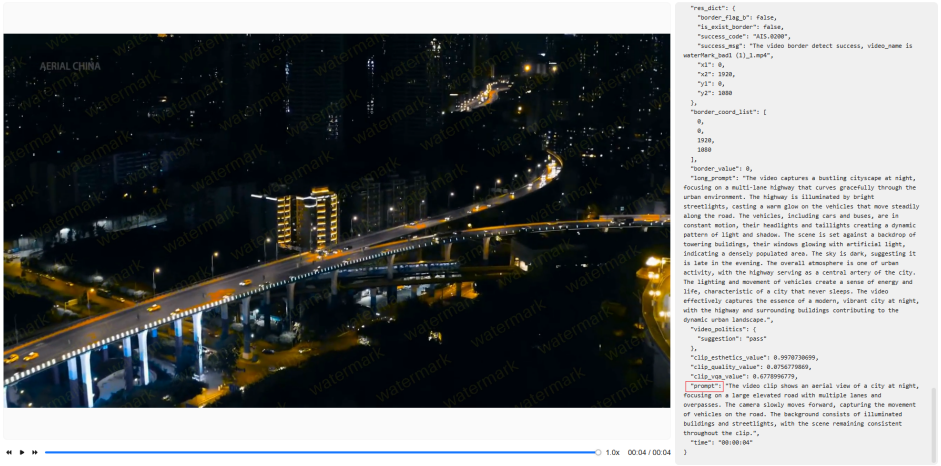
Video Synopsis Generation (Simplified)

- Applicable file format: video > mp4/avi.
- Operator description:

  Extracts frames from a video and generates a simplified video synopsis through model inference.
- Parameter configuration example:

  No parameter configuration is required.
- Example: The **prompt** field in the description indicates the simplified video synopsis.

Figure 4-70 Example



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.5.8.1.3 Image Dataset Processing Operators

The data processing operators provide multiple data operation capabilities, including data extraction, filtering, conversion, and labeling. These operators help you extract useful information from massive data and perform deep processing to generate high-quality training data.

The platform provides image and text processing operators. For details about the operator capabilities, see **Table 4-31**.

Table 4-31 Image processing operator capabilities

Category	Operator Name	Operator Description
Data extraction	**Image and text extraction**	Extracts JSON text and images from the compressed image-text package and performs structured parsing (Base64 encoding) on the images to facilitate the use of image-text processing operators.
Data filtering	**Image Metadata Filtering**	Cleans image/text data based on the image width and height, file size, and aspect ratio threshold.
	Text Length Filtering in Image-Text Pairs	Filters out the image-text pairs whose text length is not within the specified text length range. The length of a Chinese character or an English letter is counted as 1.
	Text language filtering in image-text pairs	The language type of image-text pair data is obtained through the language detection model. The image-text pair data that is not in the language to be retained will be filtered out. Note: There is a low probability that the language detection model misjudges.
	Image/text deduplication	• Deduplicates images and text based on structured images. • Checks whether the number of images corresponding to the same text exceeds the threshold. If yes, the system randomly deletes redundant images and retains only the images and text within the threshold.
	Image deduplication	Filters out duplicate image-text pairs after after image structuring.
Data labeling	**Pornographic Image Detection**	Labels image operators.
	Dangerous Situation Image Detection	Labels dangerous situation images.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Violent and Terrorism Image Detection	Filter out violent and terrorism images.
Data conversion	**Filtering of abnormal characters in images and text**	Replaces abnormal characters in the text data with null values, with data entries unchanged. • Invisible characters, for example, U+0000-U+001F • Web page labels: <p>
	Adding Watermarks	Covers the image with translucent watermark text with a fixed spacing and arranged at a specific angle.

Image and text extraction

- Applicable file formats:

  tar+jsonl: All images are saved as a TAR package. Images can be in JPG, JPEG, PNG, or BMP format. The image text is saved as a JSONL file. The image name in the JSONL file must be the same as that in the TAR package.
- Parameter description:

  Type of content to be extracted: Extract the JSON text and images from the image-text package and perform structured parsing on the images.
- Parameter configuration example:

  No parameters need to be set.

Image Metadata Filtering

- Applicable file formats:

  JPG, JPEG, PNG, and BMP

  tar: All images are saved as a TAR package. The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.
- Parameter description:

  Type of content to be filtered:

  Minimum width and height: If the width or height of an image is less than the value of this parameter, the image will be filtered out.

  Minimum file size (B): If the file size is less than the minimum file size, the file will be filtered out.

Text Length Filtering in Image-Text Pairs

- Applicable file formats:

  tar+jsonl: All images are saved as a TAR package. The images can be in JPG, JPEG, PNG, or BMP format.

  The image text is saved as a JSONL file. The image name in the JSONL file must be the same as that in the TAR package.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Parameter description:

  Type of content to be filtered: Filter out the text-text pairs whose text length is not within the text length range. The length of a Chinese character or an English letter is counted as 1.

Text Language Filtering in Image-Text Pairs

- Applicable file formats:

  tar+jsonl: All images are saved as a TAR package. The images can be in JPG, JPEG, PNG, or BMP format.

  The image text is saved as a JSONL file. The image name in the JSONL file must be the same as that in the TAR package.

- Parameter description:

  Type of content to be filtered: Extracts JSON text and images from the compressed image-text package and performs structured parsing (Base64 encoding) on the images to facilitate the use of image-text processing operators.

Image/Text Deduplication

- Applicable file formats:

  tar+jsonl: All images are saved as a TAR package. The images can be in JPG, JPEG, PNG, or BMP format.

  The image text is saved as a JSONL file. The image name in the JSONL file must be the same as that in the TAR package.

- Parameter description:

  Type of content to be filtered:

  a. Deduplicates images and text based on structured images.
  b. Checks whether the number of images corresponding to the same text exceeds the threshold. If yes, the system randomly deletes redundant images and retains only the images and text within the threshold.

Image Deduplication

- Applicable file formats:

  JPG, JPEG, PNG, and BMP

  tar: All images are saved as a TAR package. The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.

- Parameter description:

  Type of content to be filtered: After image structuring, duplicate image/text pairs are filtered out.

- Parameter configuration example:

  No parameters need to be set.

Pornographic Image Detection

- Applicable file formats:

  JPG, JPEG, PNG, and BMP

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

tar: All images are saved as a TAR package. The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.

- Parameter description:

  Type of content to be labeled: Score the pornographic content of the image. A higher score indicates a higher risk. The score range is (0, 100). Videos whose score is greater than or equal to 50 are considered pornographic videos.

- Parameter configuration example:

  No parameters need to be set.

- Detection example:

  The results are stored in the annotation file as the image_porn object.

  suggestion: indicates whether the file passes the check. **pass** indicates that the file passes the check and no problem occurs. **review** indicates that manual review is required. You can choose to bypass or block the file based on your review policy. **block** indicates that the file to be reviewed is problematic.

  confidence: detection confidence of the model. (Note that the confidence indicates the confidence of the model-provided suggestions.) If **suggestion** is **pass**, the value is 0. If **suggestion** is **review** or **block**, the value ranges from 0 to 1.

  label: label of the pornographic content detected by the model. If no pornographic content is detected, the value is empty.


```
{
  "tar_name": "data.tar",
  "file_name": "1259.jpg",
  "image_format": "jpg",
  "height": 336,
  "width": 336,
  "size": 24438,
  "base_id": "2d6dcc7c-2ee0-4928-95a8-97ece90c1134",
  "wh_radio": 1,
  "image_porn": {
    "results": [
      {
        "category": "porn",
        "confidence": 0.9956,
        "label": "sexy"
      }
    ],
    "suggestion": "block"
  },
  "filter": 0
}
```


Dangerous Situation Image Detection

- Applicable file formats:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

JPG, JPEG, PNG, and BMP

tar: All images are saved as a TAR package. The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.

- Parameter description:

  Type of content to be labeled: Labels the content of dangerous situation images.

- Parameter configuration example:

  No parameters need to be set.

- Detection example: The results are stored in the annotation file as the image_danger object.

  suggestion: indicates whether the file passes the check. **pass** indicates that the file passes the check and no problem occurs. **review** indicates that manual review is required. You can choose to bypass or block the file based on your review policy. **block** indicates that the file to be reviewed is problematic.

  confidence: detection confidence of the model. (Note that the confidence indicates the confidence of the model-provided suggestions.) If **suggestion** is **pass**, the value is 0. If **suggestion** is **review** or **block**, the value ranges from 0 to 1.

  label: label of the dangerous situation content detected by the model. If no dangerous situation content is detected, the value is empty.



```
{
  "tar_name": "data.tar",
  "file_name": "[redacted]1.png",
  "image_format": "png",
  "height": 720,
  "width": 466,
  "size": 893658,
  "base_id": "14ef093e-1ff8-4757-bc15-0963b51ca56a",
  "wh_radio": 0.6472222222,
  "image_danger": {
    "results": [
      {
        "category": "terrorism",
        "confidence": 0.9941,
        "label": "riot"
      }
    ],
    "suggestion": "block"
  },
  "filter": 0
}
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Violent and Terrorism Image Detection

- Applicable file formats:

  JPG, JPEG, PNG, and BMP

  tar: All images are saved as a TAR package. The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.

- Parameter description:

  Type of content to be labeled: Filters out violent and terrorism images.

- Parameter configuration example:

  No parameters need to be set.

- Detection example: The results are stored in the annotation file as the image_terrorism object.

  suggestion: indicates whether the file passes the check. **pass** indicates that the file passes the check and no problem occurs. **review** indicates that manual review is required. You can choose to bypass or block the file based on your review policy. **block** indicates that the file to be reviewed is problematic.

  confidence: detection confidence of the model. (Note that the confidence indicates the confidence of the model-provided suggestions.) If **suggestion** is **pass**, the value is 0. If **suggestion** is **review** or **block**, the value ranges from 0 to 1.

  label: label of the violent and terrorism content detected by the model. If no violent or terrorism content is detected, the value is empty.


```
{
  "tar_name": "data.tar",
  "file_name": "BA202310100126_6_0_Im0.png",
  "image_format": "png",
  "height": 279,
  "width": 652,
  "size": 59206,
  "base_id": "47c0346e-6dd8-47b6-89e6-1dabab4384e6",
  "wh radio": 2.3369175627,
  "image_terrorism: {
      "results": [
         {
            "category": "terrorism",
            "confidence": 0.9941,
            "label": "riot"
   }
      "suggestion": "block"
  },
  "filter": 0
}
```


 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Filtering of Abnormal Characters in Images and Text

- Applicable file formats:

  tar+jsonl: All images are saved as a TAR package. The images can be in JPG, JPEG, PNG, or BMP format.

  The image text is saved as a JSONL file. The image name in the JSONL file must be the same as that in the TAR package.
- Parameter description:

  Type of content to be converted: Replaces abnormal characters in the text data with null values, with data entries unchanged.
- Parameter configuration example:

  No parameters need to be set.

Adding Watermarks

- Applicable file formats:

  JPG, JPEG, PNG, and BMP

  tar: All images are saved as a TAR package. The images in the TAR package can be in JPG, JPEG, PNG, or BMP format.
- Parameter description:

  Watermark text: The value is of the string type.
- Conversion example

  The labeling score is stored in the element field in the JSONL file.

4.5.8.1.4 Weather Dataset Processing Operators

The platform supports the processing of weather datasets. For details about the weather processing operator capabilities, see **Table 4-32**.

Table 4-32 Weather dataset processing operator capabilities

Category	Operator Name	Operator Description
Scientific computing	**Weather data preprocessing**	Converts binary weather data files into structured JSON files.

Weather Data Preprocessing

- Applicable file formats: .nc, .cdf, .netcdf, .gr, .gr1, .grb, .grib, .grb1, .grib1, .gr2, .grb2, and .grib2
- Operator description: Converts binary weather data files into structured JSON files.
- Parameter configuration example
- Filtering example:

  After processing

{"total_size": 145364716, "geo_range": {"lon": [0.0, 360.0], "lat": [-90.0, 90.0]}, "time_range": [1641146400000, 1641146400000], "forcing_types": ["TERRAIN_HEIGHT", "SOIL_TYPE", "LAND_MASK", "OCEAN_MASK"], "surface_features": [], "surface_constants": ["TERRAIN_HEIGHT", "SOIL_TYPE", "LAND_MASK", "OCEAN_MASK"], "upper_air_layers": ["1000hPa", "100hPa", "150hPa", "175hPa", "200hPa", "250hPa", "300hPa", "400hPa", "500hPa", "50hPa", "600hPa", "700hPa", "850hPa", "925hPa"], "upper_air_features": ["Q", "T", "U", "V", "Z"], "under_sea_features": [], "under_sea_depths": []}

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.5.8.1.5 Audio Dataset Processing Operators

The platform supports the processing of audio datasets. For details about the audio processing operator capabilities, see **Table 4-33**.

Table 4-33 Audio dataset processing operator capabilities

Category	Operator Name	Operator Description
Data conversion	**Noise Addition**	Adds noise to audios.
	Noise Suppression	Removes pure noise segments from audios and reduces noise.
	Tone Change	Adjusts the pitch of the original audio.
	Reverberation Reduction	Reduces the reverberation effect of the audio in the space and improves the intelligibility of the voice.
	Voice Anonymization	Anonymizes the audio. The anonymized audio differs greatly from the the original one in the speaker timbre and voiceprint.
	Voice Noise Reduction	Reduces the noise in the original audio. Only the situation where noise and human voices overlap is considered. No restriction is imposed on pure noise audio or pure noise segments.
	Speaking Speed Adjustment	Adjusts the speaking speed in the audio.
	Voice Style Conversion	Converts the original audio based on the specified target style.
	Audio Quantization Encoding	Converts a high-resolution audio file with header information into a 16 kHz alaw/ulaw/pcm/wav file using audio encoding and decoding technologies and quantization compression technologies.
Data labeling	**Speech language recognition and labeling**	Identifies the language used by the speaker in the audio and provides the confidence.
	Speech-to-Text Conversion (Mandarin)	Converts Mandarin speech into text quickly to enrich human-machine interaction scenarios.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Operator Name	Operator Description
	Speech sentiment recognition and labeling	Recognizes the sentiments of speakers in the input audio.
	Voice Activity Detection	Detects the start and end time of each segment of human voice in the audio.
	Noise Level Evaluation	Scores the quality of the audio containing human voice segments.
	Silent Segment Detection	Identifies silent segments in the audio and the confidence, and provides the proportion of silent segments.
	Multi-speaker Speech Recognition	Identifies the audio content, and returns the start time, end time, and content of each speaker.
	Personal Privacy Dialog Identification	Labels personal privacy voice content.
	Prohibited speech detection	Labels prohibited speech.
	Political Sensitive Speech Content Recognition Operator	Labels politically sensitive speech.
	Pornographic speech content detection	Labels pornographic content.

Noise Addition

- Applicable file format: pure audio file in WAV format (audio duration ≤ 60s; sampling rate: 16 kHz)
- Operator description: Adds noise to the audio.
- Parameter description:

  Noise type: type of the noise to be added. The mixed noise is the superposition of Gaussian noise and salt-and-pepper noise.

  Signal-to-noise ratio (SNR): ratio of the normal sound signal strength to the noise signal strength.

Noise Suppression

- Applicable file format: pure audio file in WAV format (audio duration ≤ 30s; sampling rate: 16 kHz; bit depth: 16 bits; single-channel)

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Operator description: Removes pure noise segments from audios and reduces noise.
- Parameter configuration example

  No parameters need to be set.

Tone Change

- Applicable file format: pure audio file (audio duration ≤ 60s)
- Operator description: Adjusts the pitch of the original audio.
- Parameter description

  Tone: tone parameter

Reverberation Reduction

- Applicable file format: pure audio file in WAV format (audio duration ≤ 60s; sampling rate: 16 kHz)
- Operator description: Reduces the reverberation effect of the audio in the space and improves the intelligibility of the voice.
- Parameter configuration example

  No parameters need to be set.

Voice Anonymization

- Applicable file format: pure audio file in WAV format (audio duration ≤ 30s; sampling rate: 16 kHz; bit depth: 16 bits; single-channel)
- Operator description: Anonymizes the audio. The anonymized audio differs greatly from the the original one in the speaker timbre and voiceprint.
- Parameter configuration example

  No parameters need to be set.

Voice Noise Reduction

- Applicable file format: pure audio file in WAV format (sampling rate: 16 kHz; bit depth: 16 bits; single-channel)
- Operator description: Reduces the noise in the original audio. Only the situation where noise and human voices overlap is considered. No restriction is imposed on pure noise audio or pure noise segments.
- Parameter configuration example

  No parameters need to be set.

Speaking Speed Adjustment

- Applicable file format: pure audio file in WAV format (audio duration ≤ 60s)
- Operator description: Adjusts the speaking speed in the audio.
- Parameter description

  Speaking speed: The value ranges from 0.5 to 2.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Voice Style Conversion

- Applicable file format: pure audio file (file size ≤ 50 MB)
- Operator description: Converts the original audio based on the specified target style.
- Parameter description

  Voice style: voice style after conversion

Audio Quantization Encoding

- Applicable file format: pure audio file (file size ≤ 100 MB)
- Operator description: Converts a high-resolution audio file with header information into a 16 kHz alaw/ulaw/pcm/wav file using audio encoding and decoding technologies and quantization compression technologies.
- Parameter configuration example

  No parameters need to be set.

Speech Language Recognition and Labeling

- Applicable file format: pure audio file in WAV format (audio duration ≤ 60s; sampling rate: 16 kHz; bit depth: 16 bits)
- Operator description: Identifies the language used by the speaker in the audio and provides the confidence.
- Parameter configuration example

  No parameters need to be set.

Speech-to-Text Conversion (Mandarin)

- Applicable file format: pure audio file (audio duration ≤ 60s)
- Operator description: Converts Mandarin speech into text quickly to enrich human-machine interaction scenarios.
- Parameter description

  Punctuation: whether to add punctuation marks to the recognition result

  Digit conversion: whether to recognize numbers in speech as Arabic numerals

  Word segmentation information: whether the recognition result contains the word segmentation result

Speech Sentiment Recognition and Labeling

- Applicable file format: pure audio file in WAV format (audio duration ≤ 60s; sampling rate: 16 kHz; bit depth: 16 bits)
- Operator description: Recognizes the sentiments of speakers in the input audio.
- Parameter configuration example

  No parameters need to be set.

Voice Activity Detection

- Applicable file format: pure audio file in WAV format (audio duration ≤ 600s; sampling rate: 16 kHz; bit depth: 16 bits)

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Operator description: Detects the start and end time of each segment of human voice in the audio.
- Parameter configuration example

  No parameters need to be set.

Noise Level Evaluation

- Applicable file format: pure audio file in WAV format (audio duration ≤ 60s; sampling rate: 16 kHz; bit depth: 16 bits)
- Operator description: Scores the quality of the audio containing human voice segments.
- Parameter configuration example

  No parameters need to be set.

Silent Segment Detection

- Applicable file format: pure audio file (audio duration ≤ 600s; sampling rate: 16 kHz; bit depth: 16 bits)
- Operator description: Identifies silent segments in the audio and the confidence, and provides the proportion of silent segments.
- Parameter configuration example

  No parameters need to be set.

Multi-speaker Speech Recognition

- Applicable file formats: pure audio file (audio duration ≤ 1 hour; single-channel)
- Operator description: Identifies the audio content, and returns the start time, end time, and content of each speaker.
- Parameter description

  Punctuation: whether to add punctuation marks to the recognition result

  Digit conversion: whether to recognize numbers in speech as Arabic numerals

  Word segmentation information: whether the recognition result contains the word segmentation result

  Speaker separation: whether the recognition result contains speaker information

  Speaking speed: whether the recognition result contains the speaking speed

Personal Privacy Dialog Identification

- Applicable file format: pure audio file (audio duration ≤ 60s)
- Operator description: Labels personal privacy speech content.
- Parameter configuration example

  No parameters need to be set.

Prohibited Speech Detection

- Applicable file format: pure audio file (audio duration ≤ 60s)

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Operator description: Labels prohibited speech content.
- Parameter configuration example

  No parameters need to be set.

Political Sensitive Speech Content Recognition Operator

- Applicable file format: pure audio file (audio duration ≤ 60s)
- Operator description: Tags politically sensitive speech content.
- Parameter configuration example

  No parameters need to be set.

Pornographic Speech Content Moderation Operator

- Applicable file format: pure audio file (audio duration ≤ 60s)
- Operator description: Tags pornographic speech content.
- Parameter configuration example

  No parameters need to be set.

4.5.8.2 Custom Data Processing Operators

In addition to the preset processing operators, ModelArts Studio allows you to create custom processing operators to meet specific data processing requirements and service scenarios. You can flexibly define the processing logic based on the site requirements to further improve the model training effect and adaptability.

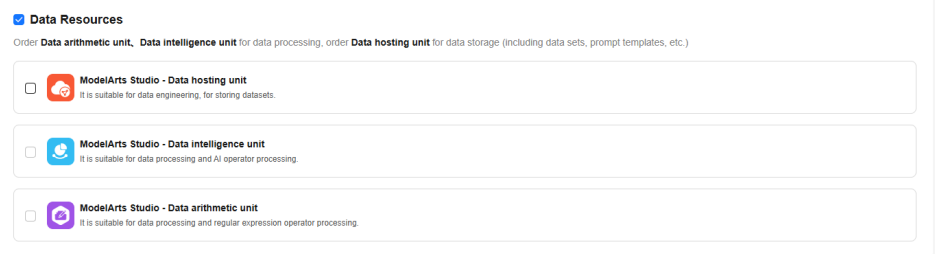
Constraints

This function is available only to yearly/monthly subscribers.

Figure 4-71 Applying for trial



Figure 4-72 Yearly/Monthly resources



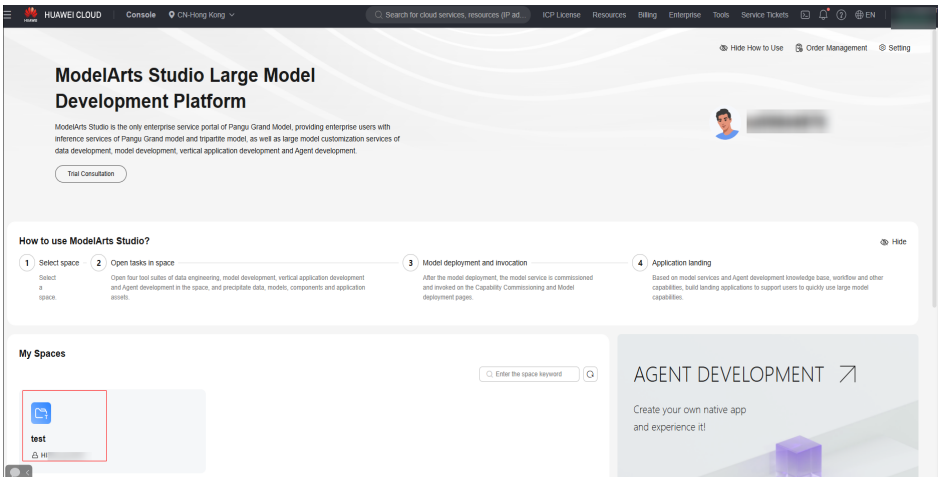
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Creating a Custom Processing Operator

To create a custom processing operator, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

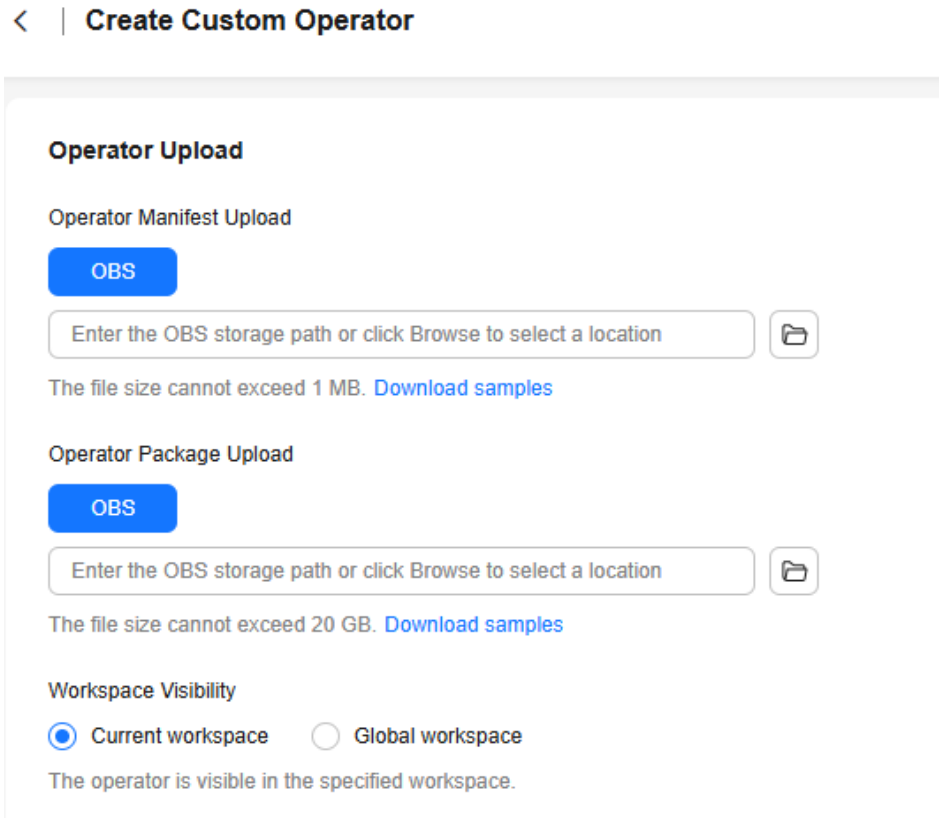
   Figure 4-73 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Processing** > **Processing Tasks**. Click **Manage Processing Operator** in the upper right corner.
3. On the **Manage Processing Operators** page, click the **Custom** tab, and click **Create Custom Operator** in the upper right corner.
4. On the **Create Custom Operator** page, click **Download samples** to view the specifications of the operator configuration file and operator package. Use OBS to upload the operator configuration file and operator package, set the workspace visibility, and click **OK** in the lower right corner.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-74 Creating a custom operator



5. Custom operators that are created can be used in **Processing Text Datasets**, **Processing Image Datasets**, **Processing Video Datasets**, **Processing Weather Datasets**, and **Processing Other Datasets**.

Viewing operator details:

Click the operator name to view its details.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Operator Configuration File Specifications

Table 4-34 Basic information configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
id	string	Yes	Operator name in English.	The operator name must start with a letter and can contain up to 128 characters, including letters, digits, and underscores (_). If the length exceeds 128 characters, the excess part is automatically truncated. The ID cannot be changed when the operator is updated. The ID of a visible operator in all workspaces is unique in all workspaces, and the ID of a visible operator in the current workspace is unique in the current workspace.
name	string	Yes	Operator display name.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
description	string	No	Operator description.	The length cannot exceed 2000 characters. The excess part is automatically truncated.
author	string	No	Developer name.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
tags	**tags** object	Yes	Operator label, which is used for classification and filtering.	For details, see **Table 4-35**.
labels	Array of **label** objects	No	Label output by the labeling operator.	For details, see **Table 4-40**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Type	Mandatory	Description	Constraints
runtime	**runtime** object	Yes	Operator running configuration.	For details, see **Table 4-36**.
arguments	Array of **argument** objects	No	List of operator input parameters.	For details, see **Table 4-38**.

Table 4-35 Tag configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
language	Array of strings	Yes	List of languages that can be processed by the operator, for example, zh and en.	The length of a single character string cannot exceed 32 characters. The excess part is automatically truncated. Enter an international language code.
format	Array of strings	Yes	List of dataset file name extensions supported by the operator, for example, JSON, CSV, and MP4.	The length of a single character string cannot exceed 32 characters. The excess part is automatically truncated.
category	string	Yes	Operator type.	Select only one option. The options are as follows: - Data extraction - Data sampling - Data conversion - Data filtering - Data deduplication - Data labeling - Other

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Type	Mandatory	Description	Constraints
modal	Array of strings	Yes	List of data modalities supported by the operator.	Select one or multiple options. The options are as follows: - TEXT - IMAGE - VIDEO - AUDIO - OTHER (including weather and prediction)
custom	Array of strings	No	List of custom operator tags.	The length of a single character string cannot exceed 32 characters. The excess part is automatically truncated.

Table 4-36 runtime configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
cpu-arch	Array of strings	Yes	List of CPU architectures supported by the operator.	Select one or multiple options. The options are as follows: - Arm - x86
xpu-devices	Array of strings	No	List of device models supported by the operator. This parameter is mandatory when the operator runs on NPUs.	The options are as follows: - SNT9B This parameter is mandatory when the number of NPUs of a resource in runtime.resources is greater than 0. The value can be SNT9B.
environment	string	Yes	Operator package type. Only pure Python operator packages are supported.	Select only one option. The options are as follows: - PYTHON

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Type	Mandatory	Description	Constraints
entrypoint	string	Yes	Operator startup command.	The value contains a maximum of 128 characters. The excess part is automatically truncated. If environment is set to python, enter the fixed name process.py. The name cannot be changed.
auto-data-loading	boolean	Yes	Whether to automatically load data.	If the value is true, the framework processes the input and output. If the value is false, the user processes the input and output.
resources	Array of **resource** objects	No	List of resource sizes required for running a single operator instance.	This parameter is mandatory when environment is set to python.

Table 4-37 resource configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
cpu	integer	Yes	Default CPU resources of a single instance (unit: vCPU).	You need to configure it based on the actual number of general computing units.
memory	integer	Yes	Default memory size of a single instance, in MB.	You need to configure it based on the actual number of general computing units.
npu	integer	No	Default number of NPUs of a single instance, in card.	You need to set this parameter based on the specifications and quantity of the subscribed intelligent computing units.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-38 argument configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
key	string	Yes	Parameter ID, which must be unique in the parameter list.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
name	string	Yes	Parameter display name.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
type	string	Yes	Parameter type.	Select only one option. The options are as follows: - STRING - FLOAT - INT - ENUM (radio button) - LIST (check box) - OBS (front-end component parameter for selecting an OBS file) - BOOLEAN
tips	string	No	Parameter description.	The length cannot exceed 2000 characters. The excess part is automatically truncated.
min	float	No	Minimum value of the parameter. This parameter is optional when type is set to INT or FLOAT.	A maximum of four decimal places are reserved for floating-point numbers.
max	float	No	Minimum value of the parameter. This parameter is optional when type is set to INT or FLOAT.	A maximum of four decimal places are reserved for floating-point numbers.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Type	Mandatory	Description	Constraints
between	boolean	No	Whether the parameter is a range-type numeric parameter. This parameter is optional when type is set to INT or FLOAT. The default value is false.	-
items	Array of **item** objects	No	List of enumerated values. This parameter is mandatory when type is set to ENUM or LIST.	This parameter is mandatory when type is set to ENUM or LIST. The number of enumerated items must be at least 1.
required	boolean	No	Whether the parameter is mandatory.	Set it to **true** or **false**.
visible	boolean	No	Whether the parameter is visible to the frontend.	Set it to **true** or **false**.
default	string	No	Default parameter value.	If **visible** is set to **false** and **requires** is set to **true**, the default value must be set. If there are multiple default values, separate them with commas (,), for example, SD,HD. The default value of a numeric parameter is in the format of min;max.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 4-39 item configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
name	string	Yes	Enumerated item value.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
value	string	Yes	Value of the enumerated item.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
tips	string	No	Description or description of the enumerated item.	The length cannot exceed 2000 characters. The excess part is automatically truncated.

Table 4-40 label configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
key	string	Yes	Key of a tag.	The key can contain a maximum of 128 characters. If the key exceeds 128 characters, the excess part is automatically truncated. The key must be unique among operators.
name	string	Yes	Tag name.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
type	string	Yes	Tag type.	Enumerated value. The options are STRING, NUMERIC, ENUM, and OBJECT. Only one option can be selected.
min	float	No	Minimum tag value.	This parameter is mandatory when type is set to NUMERIC.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Type	Mandatory	Description	Constraints
max	float	No	Maximum tag value.	This parameter is mandatory when type is set to NUMERIC.
items	Array of **labelItem** objects	No	Tag enumeration list.	This parameter is mandatory when type is set to ENUM.
dimensions	Array of **labelDimension** objects	No	Level-1 tag dimension.	This parameter is mandatory when type is set to OBJECT.

Table 4-41 labelItem configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
name	string	Yes	Enumerated item name.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
value	string	Yes	Enumerated value.	The value contains a maximum of 128 characters. The excess part is automatically truncated.

Table 4-42 labelDimension configuration specifications

Parameter Name	Type	Mandatory	Description	Constraints
key	string	Yes	Key of the level-2 tag.	The value contains a maximum of 128 characters. The excess part is automatically truncated.
name	string	Yes	Level-2 tag name.	The value contains a maximum of 128 characters. The excess part is automatically truncated.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter Name	Type	Mandatory	Description	Constraints
type	string	Yes	Level-2 tag type.	Enumerated value. The options are STRING, NUMERIC, and ENUM. Only one option can be selected.
min	float	No	Minimum value of the level-2 tag.	This parameter is mandatory when type is set to NUMERIC.
max	float	No	Maximum value of the level-2 tag.	This parameter is mandatory when type is set to NUMERIC.
items	Array of **labelItem** objects	No	Enumeration list of level-2 tags.	This parameter is mandatory when type is set to ENUM.

An operator configuration file is a YAML file that describes the basic information, function parameters, operating environment, and resource requirements of an operator. Configure the configuration file as follows:

```
id: video_clip    # (Mandatory) Operator abbreviation. The value must start with a letter and can contain a maximum of 128 characters, including letters, digits, and underscores (_). If the value contains more than 128 characters, the excess characters will be truncated.
name: Video clip # (Mandatory) Operator display name. The value can contain a maximum of 128 characters. If the value contains more than 128 characters, the excess characters will be truncated.
description: Splits a long video into multiple clips. # (Mandatory) Operator description. The value can contain a maximum of 2,000 characters.
author: "xxx Technology Co., Ltd." # (Optional) Developer or team name, for example, Data Team. The value can contain a maximum of 128 characters.
tags:  # (Mandatory) Operator tag, which is used for classification and filtering in specific scenarios.
 language: # (Mandatory) Language that can be processed by the operator. The value is a language code. Multiple options can be selected. Only international standard language codes are supported. The value contains a maximum of 32 characters.
   - zh
   - en
   -...
 format: # (Mandatory) Dataset file format supported by the operator. Multiple options are supported. The value contains a maximum of 32 characters.
     - JSONL
     - TXT
     - CSV
     - HTML
     - MOBI
     - EPUB
     - DOCX
     - PDF
     - MP4
     - AVI
     -...
 category: # (Mandatory) Operator type, which is used to display operators by category on the GUI. Only one option can be selected. The options are as follows:
   - Data extraction
   - Data sampling
   - Data conversion
   - Data filtering
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
  - Data Deduplication
  - Data labeling
  - Other
 modal:  # (Mandatory) Data modality supported by the operator. Multiple options can be selected. The options are as follows:
  - TEXT
  - IMAGE
  - VIDEO
  - AUDIO
  - OTHER
 custom: # Tag of a custom operator. Multiple options can be selected. The value contains a maximum of 32 characters.
  - Data augmentation
  - Pre-labeling
  -...
runtime:
 cpu-arch: # (Mandatory) Supported CPUs
  - ARM
  - X86
 xpu-devices: # XPU model supported by the operator. Multiple options can be selected. The options are as follows (mandatory if NPU is involved):

  - SNT9B

 resources: # Default resources of a single instance. This parameter is mandatory when the operator package type is python.
  - cpu: 16 # CPU processor type
    memory: 256
  - cpu: 8 # NPU processor type
  - memory: 1024 # Unit: MB
    npu: 1

 environment: python # (Mandatory) Operator package type. The value can be python (pure Python operator package).
 entrypoint: process.py # (Mandatory) Fixed file name process.py when environment is set to python. The value cannot be changed.
#All service parameters are input parameters of the operator. The data types of the parameters include STRING, FLOAT, INT, ENUM (radio button), LIST (check box), and OBS.
arguments:
 # Example of parameters of the STRING type
 - key: filter_keywords # [Mandatory]
   name: Filter keyword.
   type: STRING
   tips: The samples that match the keywords will be filtered. Multiple keywords are separated by commas (,).  #Tips on the GUI
   required: true
   visible: true
   default: gambling  # Default value of the parameter. This parameter is optional. Use commas (,) to separate multiple default values.

 # Example of the INT/FLOAT value range type
 - key: length_of_characters # [Mandatory]
   name: filtering duration range
   type: FLOAT # [Mandatory] When type is set to int or float, the between field is mandatory.
   between: true # Whether to obtain the value range. When type is set to int or float, the between field is mandatory.
  min: 1.0  # Minimum value of the parameter, which is optional. When type is set to float, the value can contain a maximum of four decimal places. If the value exceeds the maximum, it will be truncated.
   max: 500.0   # Range of the maximum parameter value. This field is optional.
   tips: The filtering duration range includes the entered boundary value. The value is of the float type, in characters. Samples whose text length is within the specified range are retained. #GUI tips
   required: true
   visible: true
   default: 100.0;300.0 # (Optional) Default maximum and minimum parameter values. If this parameter is not set, min;max is used as the default value.

 # Example of parameters of the INT/FLOAT numeric type
 - key: max_cropping_area_ratio
   name: Maximum cropping area ratio
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
    type: FLOAT
    between: false # Whether the parameter is of the range type. The default value is false.
    min: 0.0   # Range of the minimum value of the parameter
    max: 100.0    # Range of the maximum value of the parameter
    tips: The value is a float ranging from 0.0 to 100.0, in percentage (%). Samples whose cropped area ratio
(cropped video area/original video area) is greater than the value will be filtered out.
    visible: true
    required: true
    default: 100 # Default value

  # Example of an ENUM parameter
  - key: font_conversion
    name: Text font conversion
    type: ENUM # Only one option can be selected.
    items:
      - name: Simplified Chinese to Traditional Chinese
        value: traditional
      - name: Convert Traditional Chinese to Simplified Chinese
        value: simplified
    required: true
    visible: true
    default: simplified

  # Example of a LIST parameter
  - key: resolution
    name: Resolution
    type: LIST # Multiple options can be selected.
    items: # Parameter options. This parameter is mandatory when type is set to ENUM or LIST.
      - name: Smoothness   #[Mandatory]
        value: SM   #[Mandatory]
        tips: 480 > resolution ≥ 360
        - name: SD
        value: SD
        tips: 720 > resolution ≥ 480
      - name: HD
        value: HD
        tips: 1080 > resolution ≥ 720
    required: true
    visible: true
        default: SD,HD # Use commas (,) to separate multiple default values.

  # Example of an OBS parameter
  - key: sensitive_word
    name: OBS path of the sensitive word dictionary file
    type: OBS
    tips: sensitive word dictionary file
    required: true
    visible: true
    default: NLP/system_resource/sensitive_word.csv # OBS path of the default word dictionary
  # Example of a BOOLEAN parameter
  - key: parse_all
    name: Whether to parse all files
    type: BOOLEAN
    items:
      - name: Yes
        value: true
      - name: No
        value: false
    visible: true
    required: true
    default: false
```

Operator Package Specifications

Python operator package

Assume that the operator package name is **video_clip.tar**. The directory structure after the operator package is decompressed is as follows:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
+--- video_clip # The directory name must be the same as the tar package name.
| +--- program_package # Python operator directory
| | +--- install.sh # (Optional) Installation script
| | +--- process.py # (Mandatory) Operator code
```

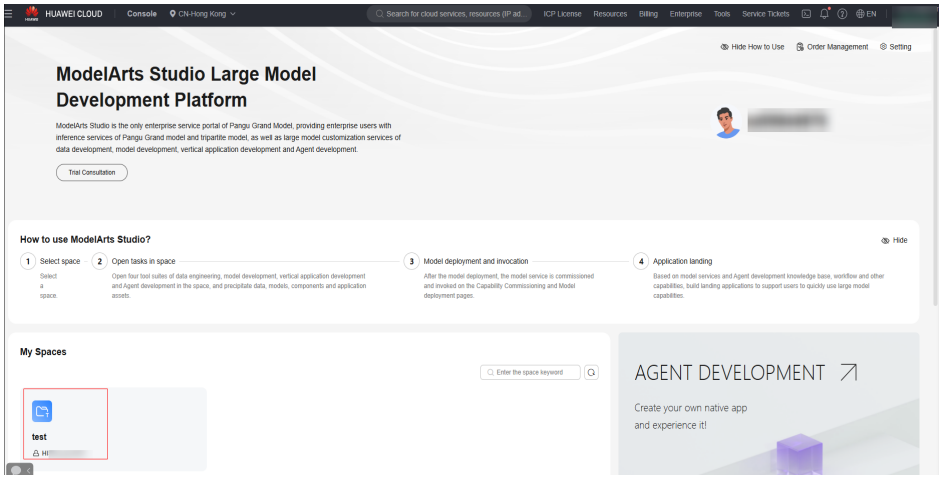
4.5.9 Managing Processed Datasets

If you have completed data processing, data synthesis, data labeling, or data combination tasks for a dataset, click **Generate** in the corresponding task list to generate a processed dataset. The generated datasets are managed by the platform in a unified manner and used for subsequent publishing tasks.

The platform allows you to view basic information and data lineage of processed datasets. The procedure is as follows:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-75 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Datasets**. Click the **Processed Dataset** tab.
3. Click the dataset name to view the basic information, data preview, data lineage, and operation records of the processed dataset.
   - On the **Basic Information** tab page, you can view the data details, data source, and extended information. In the **Extended Info** area, you can set dataset properties as required, including the dataset property name, industry, language, and custom tag.
   - On the **Data Preview** tab page, you can view the processed data.
   - On the **Data Lineage** tab page, view the operations performed on the dataset, such as import and synthesis.
   - On the **Operation Record** tab page, view the operations and status of the dataset.
4. Click **Delete** in the **Operation** column to delete unnecessary datasets.
   - To restore a deleted dataset, click **Show Deleted Data** in the upper right corner. In the list of deleted datasets, restore the dataset.
   - To permanently delete a dataset, click the dataset name to go to the dataset details page, confirm the dataset content, and delete it permanently.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

NOTE

Deleting a processed dataset is a high-risk operation. Before deleting a dataset, ensure that it is no longer used.

4.5.10 Managing Processing Task Resources

When creating MRS-related processing or publishing tasks, you can set task resource parameters. For details about the parameters, see **Table 4-43**.

Table 4-43 Task resource parameters

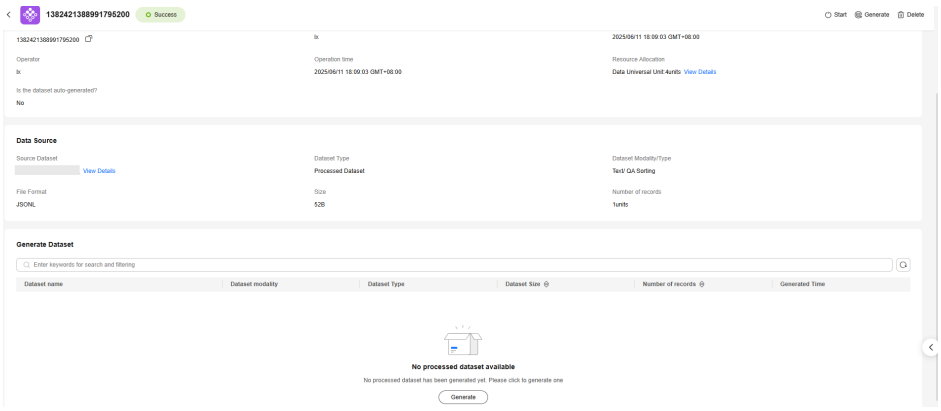
Parameter Name	Description
executorCores	Number of CPU cores of each Executor in a Spark application.
numExecutors	Number of Executors in a Spark application.
executorMemory	Executor memory of a Spark application, for example, 2 GB or 2048 MB.
driverCores	Number of CPU cores of the Spark application driver. This configuration will replace the default setting in **sc_type**.
driverMemory	Driver memory of a Spark application, for example, 2 GB or 2048 MB.

4.5.11 Generating a Dataset in a Processing Task

The dataset generation function is provided on the processing task details page.

1. When a processing task is complete and no dataset is generated, click the **Generate** button to generate a dataset if needed.

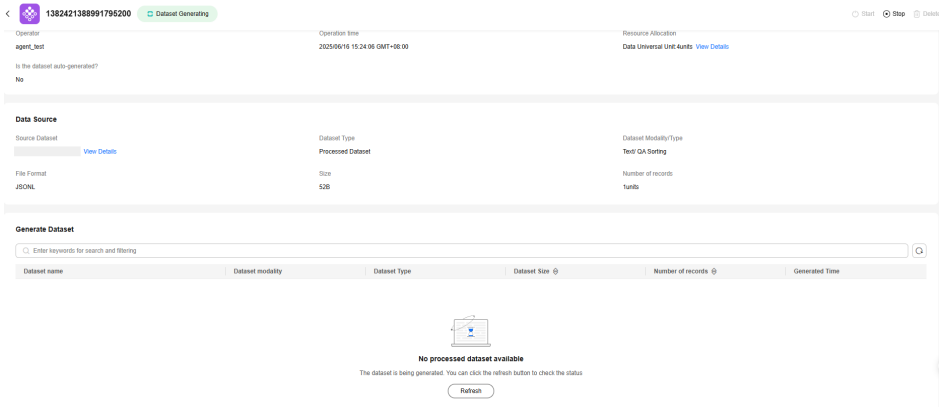
Figure 4-76 Generate button



2. When no dataset is generated for the task and you need to view the status, click the **Refresh** button to refresh the table.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-77 Refresh button



3. When the latest processing task is successfully executed and no latest dataset is generated, the message "The task data has been updated. To save the latest result, click Generate Processed Dataset." is displayed.

4.6 Publishing a Dataset

4.6.1 Dataset Publishing Scenarios

Introduction to Dataset Publishing

ModelArts Studio provides data evaluation and data publishing functions. It aims to ensure that data meets the diversity, balance, and representativeness requirements of large model training through data quality evaluation, promoting efficient data circulation and application.

Data publishing includes not only publishing data in a suitable format, but also evaluating the dataset effectiveness based on task requirements to ensure that the dataset meets the standards of model training in terms of scale, quality, and content.

- **Data evaluation**

  The platform offers predefined evaluation standards for multiple types of data. You can choose from these predefined standards or customize evaluation standards as needed to precisely improve data quality, ensure that data meets high standards, and enhance model performance.

- **Data publishing**

  Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations. The following formats are supported: standard format and Pangu format (applicable to Pangu model training). Currently, only text and image datasets can be published in Pangu format.

With these functions, the platform can help you scientifically manage and publish datasets to ensure that the dataset quality meets the requirements of large model training, thereby improving the effect of subsequent model training.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Significance of Data Publishing

Data publishing allows you to convert data into different formats, and to evaluate the dataset effect based on task requirements to ensure that the data meets the training standards in terms of scale, quality, and content. Specifically, data publishing has the following significance:

- **Support for multiple formats**

  For text and image datasets, the platform supports multiple data formats, including standard format and Pangu format, to meet different training requirements. You can convert datasets into a required format to ensure that the data is compatible with specific models (such as Pangu models) and to optimize the training effect.
- **Improved training efficiency**

  Publishing datasets that comply with standards can greatly improve data processing efficiency, reduce subsequent adjustment workload, and help you quickly enter the model training phase.

Dataset publishing is a key step in data engineering to ensure that datasets meet model training requirements. With the data publishing function provided by the platform, you can flexibly select the format of published data based on specific task requirements to ensure data compatibility and consistency, laying a solid foundation for subsequent model training and application deployment.

Dataset Types That Support Data Publishing

Table 4-44 lists the dataset types that support data publishing.

Table 4-44 Dataset types that support data publishing

Data Type	Data Evaluation	Data Publishing
Text	√	√
Image	√	√
Video	√	√
Weather	-	√
Prediction	-	√
Other	-	√

ModelArts Studio allows you to publish **text and image datasets** in either of the following formats:

- **Standard format**: applies to a wide range of data application scenarios and meets most standard requirements of model training. Datasets in this format are published to assets, but are invisible to downstream model development personnel.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- **Pangu format**: a format designed for Pangu model training to ensure the compatibility and consistency of datasets in Pangu model training. Datasets in this format will be used for model development on ModelArts Studio.

Datasets other than text and image datasets can only be published as standard datasets.

4.6.2 Publishing Text Datasets

4.6.2.1 Evaluating Text Datasets

Before publishing a text dataset, ModelArts Studio allows you to evaluate the dataset to optimize data quality, ensure that the data meets high standards, and improve model performance.

If data evaluation is not required, skip this section and proceed with **Publishing Text Datasets**.

Creating Evaluation Standards for Text Datasets

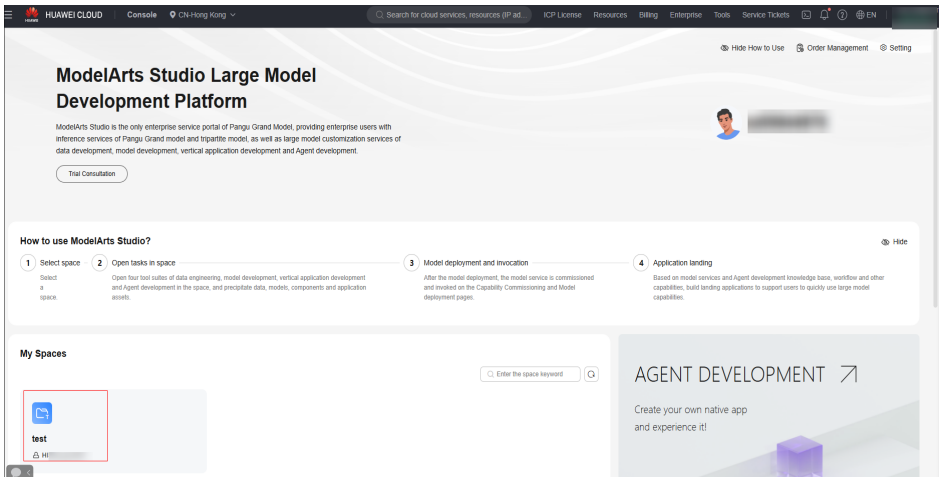
ModelArts Studio provides preset evaluation standards for text datasets, covering multiple dimensions such as data accuracy, integrity, consistency, and format specifications. You can directly use those standards or create your evaluation standards.

If you want to use the preset evaluation standards, skip this section and proceed with **Creating a Text Dataset Evaluation Task**.

To create evaluation standards for text datasets, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

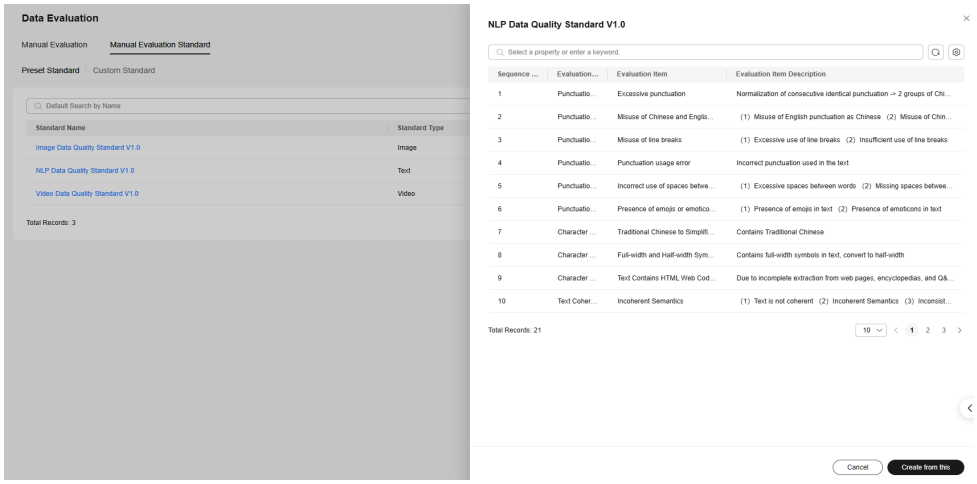
   Figure 4-78 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Data Evaluation**. On the **Manual Evaluation Standard** tab page, the preset text dataset evaluation standard **NLP Data Quality Standard V1.0** are displayed. You can click the evaluation criterion name to view the specific evaluation items.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-79 Preset evaluation standards for text datasets



3. On the **Manual Evaluation Standard** page, click **Create custom standards**, select **Preset Standard**, and set the evaluation standard name and description.
4. Edit evaluation items.

   You can delete evaluation items or create custom evaluation items as required. When creating a custom evaluation item, ensure that the evaluation type, evaluation item, and evaluation item description are clear and unambiguous.
5. Click **Complete Creation**.

   After evaluation standards are created, you can view, edit, and delete them on the **Manual Evaluation Standard** page.

Creating a Text Dataset Evaluation Task

Only processed datasets can be evaluated.

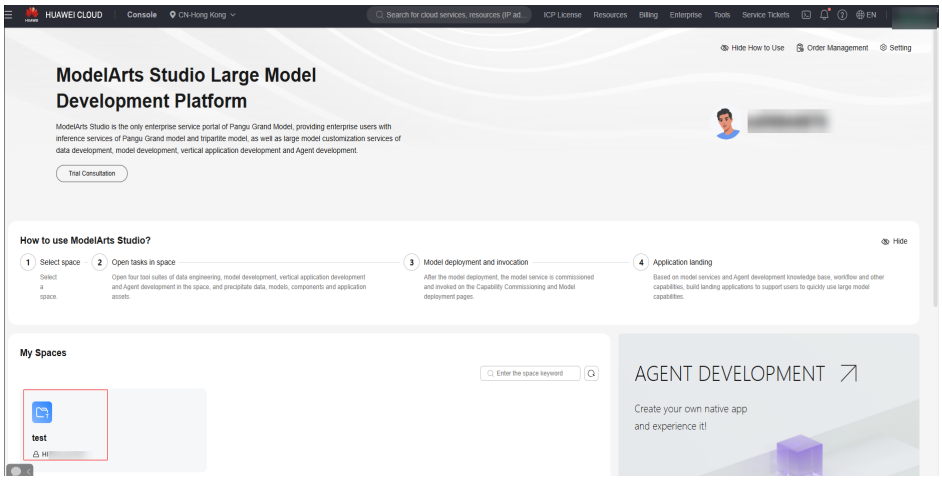
Before creating a text dataset evaluation task, generate a processed dataset by referring to **Processing Text Datasets**.

To create a text dataset evaluation task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

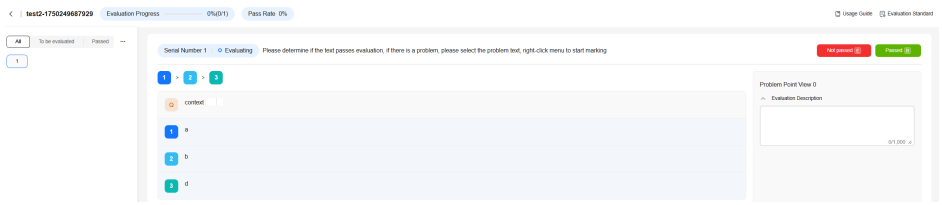
Figure 4-80 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Data Evaluation**. On the displayed page, click **Create Evaluation Task** in the upper right corner.
3. Select the dataset to be evaluated and set the number of samples.
4. Click **Next Step** and select the evaluation standard. Click **Next Step** to set the evaluator and click **Next Step** to enter the task name.
5. Click **Complete Creation** to go to the **Data Evaluation** page. After the evaluation task is created, the status becomes **Created**.
6. Click **Evaluate** in the **Operation** column.
7. On the evaluation page, label the problems of the current data by referring to the evaluation items. If the data meets the requirements, click **Passed**. If the data does not meet the requirements, click **Not passed**.

   For a text dataset, you can right-click the question content and choose the corresponding problem type from the shortcut menu as shown in **Figure 4-81**.

Figure 4-81 Marking dataset problems



8. After all data is evaluated, check that the evaluation progress is 100% on the **Manual Evaluation** page.

   Click **Report** in the **Operation** column to view the dataset quality evaluation report.

4.6.2.2 Publishing Text Datasets

Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations.

Text datasets can be published in the following formats:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- **Standard format**: original format supported by the data project function.

  The following is an example of the standard format, where **context** and **target** are key-value pairs.

  ```
  {"context": "Hello, please introduce yourself.","target": "I am a Pangu model."}
  ```

- Pangu format: To train a Pangu model, publish the dataset in Pangu format.

  The following is an example of the Pangu format, where **context** and **target** are key-value pairs. Different from the standard format, **context** is an array.

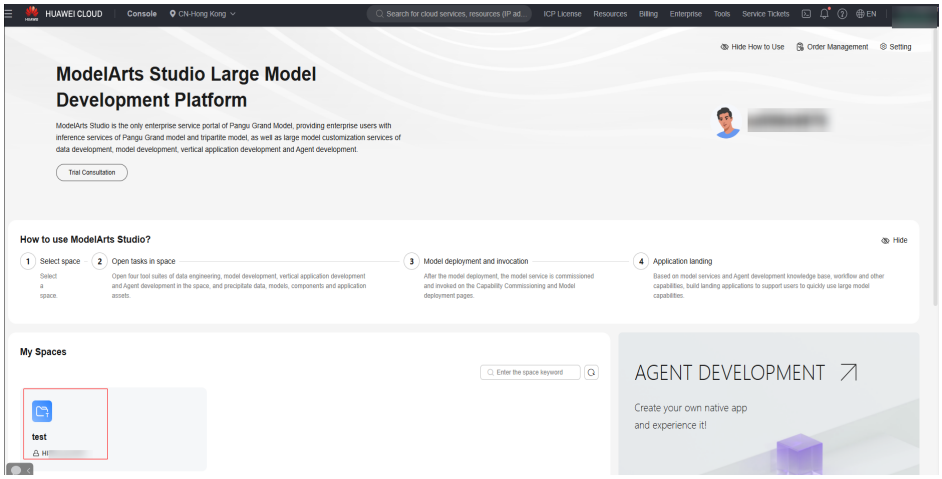
  ```
  {"context": "Hello, please introduce yourself.","target": "I am a Pangu model."}
  ```

Creating a Text Dataset Publishing Task

To create a text dataset publishing task, perform the following steps:

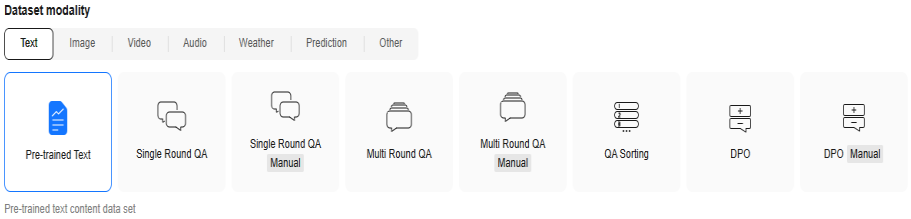
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-82 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
3. On the **Create Data Publish Task** page, select a dataset modality, for example, **Text** > **Pre-trained Text**.

   Figure 4-83 Selecting dataset modality

4. Select a dataset and click **Next**.
5. In the **Basic Configuration** area, select the data usage, dataset visibility, and application scenario.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Data engineering supports the interconnection with Pangu models. To ensure that these datasets can be properly trained by these large models, the platform supports the publishing of datasets in different formats.

Currently, the standard and Pangu formats are supported.

- **Standard format**: original format supported by the data project function. Datasets in this format can be published to assets but are not visible to downstream model development. This format is used when **Application Scenario** is set to **Other**.
- **Pangu format**: data format required for Pangu model training. The dataset will be used for model development on ModelArts Studio. This format is used when **Application Scenario** is set to **STUDIO Model Training**.

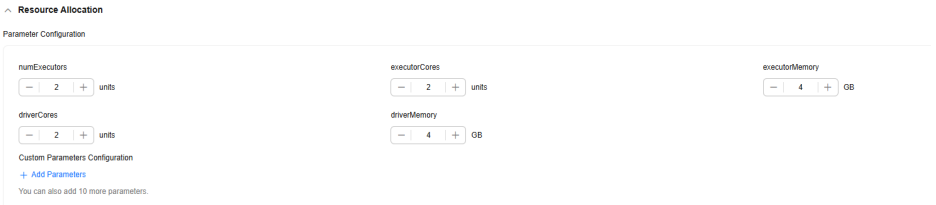
NOTE

If this dataset is used to train Pangu models, set **Select Format** to **Pangu format**.

6. Enter the dataset name and description, set extended information, and click **Next**.
7. On the **Task Configuration** page, configure resources.

   Expand **Resource Allocation** to configure task resources. You can also customize parameters. Click **Add Parameters** and enter the parameter name and value.

   Figure 4-84 Resource Allocation



8. After the task is configured, click **OK** to publish the dataset. If the task status is **Succeeded**, the data publishing task is successfully executed. You can choose **Data Engineering** > **Data Publishing** > **Datasets** in the navigation pane and click the **Published Dataset** tab to view the published dataset.

4.6.3 Publishing Image Datasets

4.6.3.1 Evaluating Image Datasets

Before publishing an image dataset, ModelArts Studio allows you to evaluate the dataset to optimize data quality, ensure that the data meets high standards, and improve model performance.

If data evaluation is not required, skip this section and proceed with **Publishing Image Datasets**.

Creating Evaluation Standards for Image Datasets

ModelArts Studio provides preset evaluation standards for image datasets, covering multiple quality dimensions such as image definition, resolution, label

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

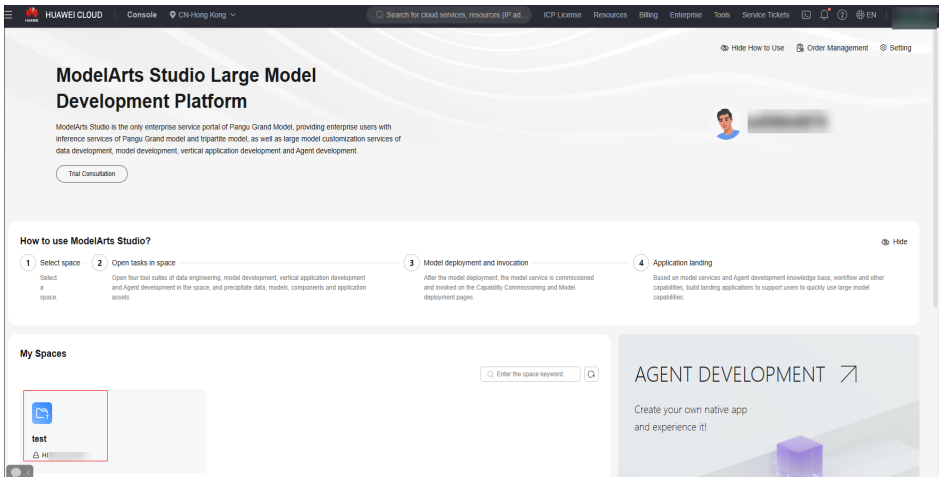
accuracy, and image consistency. You can directly use those standards or create your evaluation standards.

If you want to use the preset evaluation standards, skip this section and proceed with **Creating an Image Dataset Evaluation Task**.

To create evaluation standards for image datasets, perform the following steps:

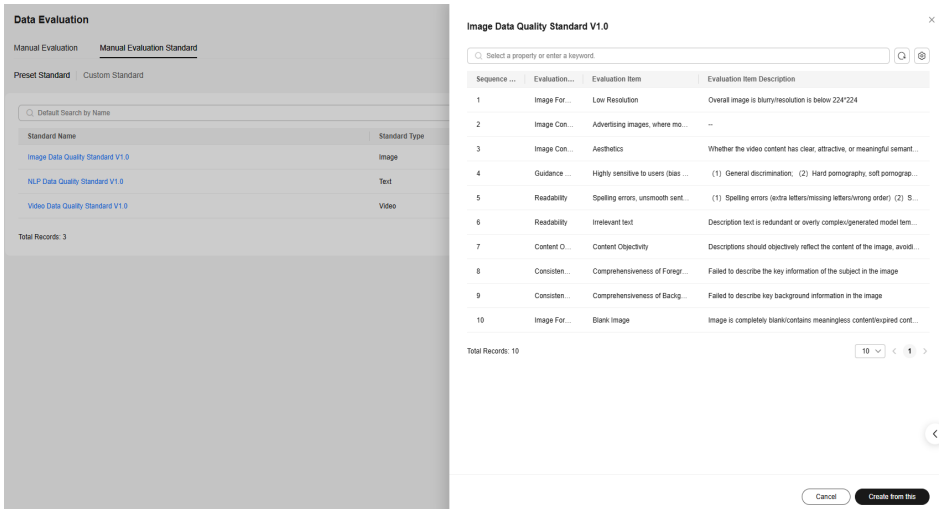
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 4-85 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Data Evaluation**. On the **Manual Evaluation Standard** tab page, the preset image dataset evaluation standard **Image Data Quality Standard V1.0** are displayed. You can click the evaluation standard name to view the specific evaluation items.

Figure 4-86 Preset evaluation standards for image datasets



3. On the **Manual Evaluation Standard** page, click **Create custom standards**, select **Preset Standard**, and set the evaluation standard name and description.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4. Edit evaluation items.

   You can delete evaluation items or create custom evaluation items as required. When creating a custom evaluation item, ensure that the evaluation type, evaluation item, and evaluation item description are clear and unambiguous.

5. Click **Complete Creation**.

   After evaluation standards are created, you can view, edit, and delete them on the **Manual Evaluation Standard** page.

Creating an Image Dataset Evaluation Task

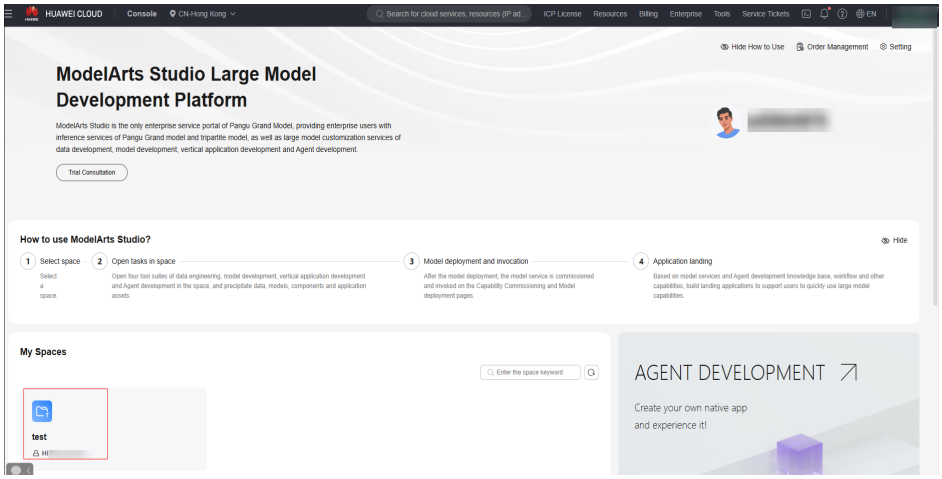
Only processed datasets can be evaluated.

Before creating an image dataset evaluation task, generate a processed dataset by referring to **Processing Image Datasets**.

To create an image dataset evaluation task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-87 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Data Evaluation**. On the displayed page, click **Create Evaluation Task** in the upper right corner.
3. Select the dataset to be evaluated and set the number of samples.
4. Click **Next Step** and select the evaluation standard. Click **Next Step** to select the evaluator, and click **Next Step** to enter the task name.
5. Click **Complete Creation** to go to the **Data Evaluation** page. After the evaluation task is created, the status becomes **Created**.
6. Click **Evaluate** in the **Operation** column.
7. On the evaluation page, label the problems of the current data by referring to the evaluation items. If the data meets the requirements, click **Passed**. If the data does not meet the requirements, click **Not passed**.
8. After all data is evaluated, check that the evaluation progress is 100% on the **Manual Evaluation** page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Click **Report** in the **Operation** column to view the dataset quality evaluation report.

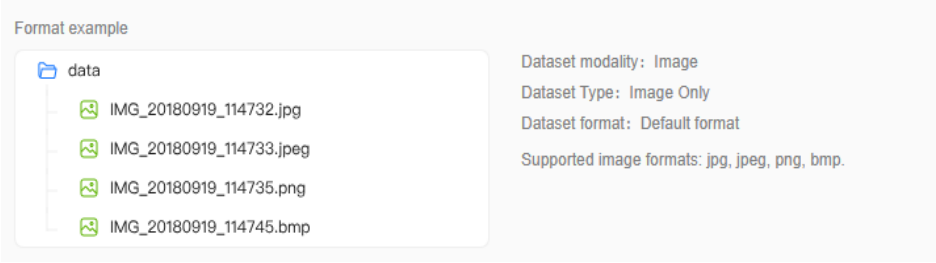
4.6.3.2 Publishing Image Datasets

Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations.

Image datasets can be published in the following formats:

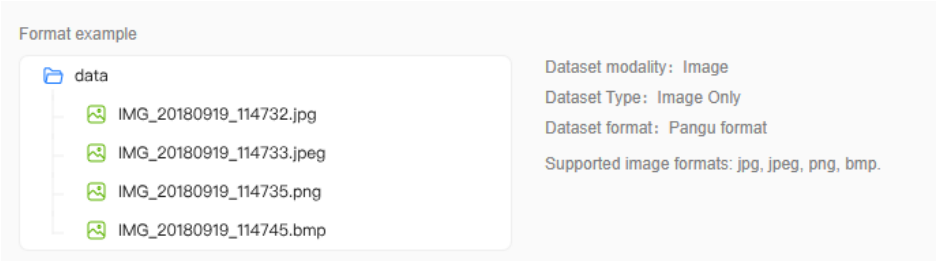
- **Standard format:** It is the default dataset format on the platform, as shown in **Figure 4-88**. Datasets in this format can be published to assets but are not visible to downstream model development.

  Figure 4-88 Example of an image dataset in Standard format

- **Pangu format:** To train a Pangu model, publish the dataset in Pangu format, as shown in **Figure 4-89**. The dataset will be used for model development on the ModelArts Studio Large Model Development Platform.

  Figure 4-89 Example of an image dataset in Pangu format

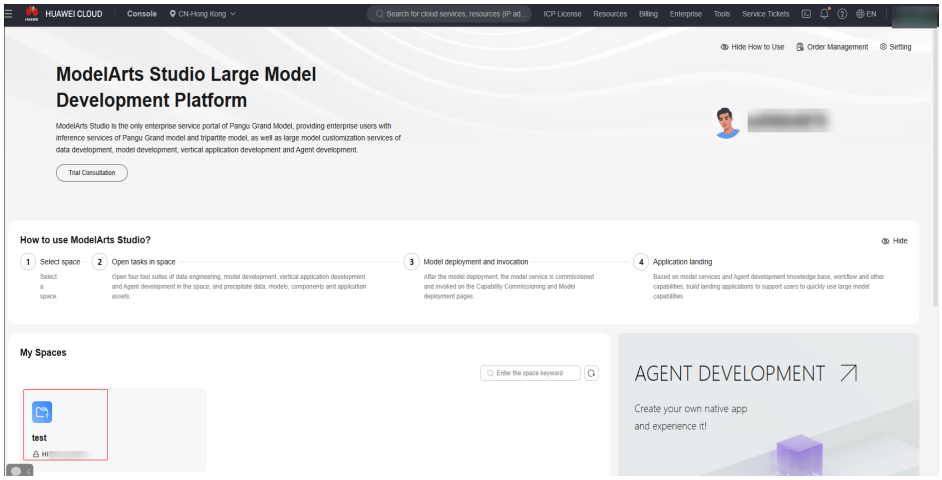
Creating an Image Dataset Publishing Task

To create an image dataset publishing task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

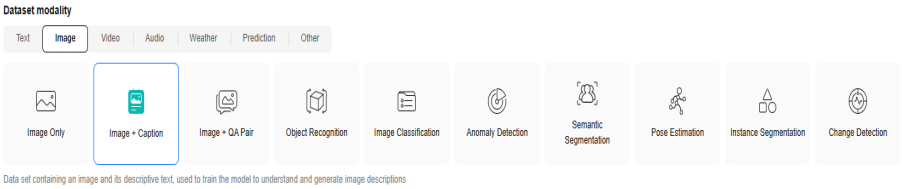
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-90 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
3. On the **Create Data Publish Task** page, select a dataset modality, for example, **Image** > **Image + Caption**.

Figure 4-91 Selecting dataset modality



4. Select a dataset and click **Next**.
5. In the **Basic Configuration** area, select the data usage, dataset visibility, and application scenario.

   Data engineering supports the interconnection with Pangu models. To ensure that these datasets can be properly trained by these large models, the platform supports the publishing of datasets in different formats.

   Currently, the standard and Pangu formats are supported.

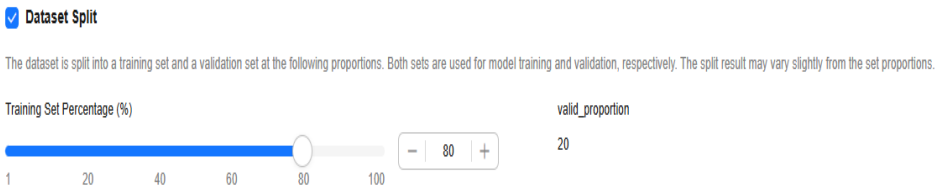
   - **Standard format**: original format supported by the data project function. Datasets in this format can be published to assets but are not visible to downstream model development.
   - **Pangu format**: data format required for Pangu model training. The dataset will be used for model development on ModelArts Studio.

   NOTE

   If this dataset is used to train Pangu models, set **Format Configuration** to **Pangu format**.
6. Select **Dataset splitting** if required. If it is selected, set the data ratio of the training set to the validation set, as shown in **Figure 4-92**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-92 Dataset splitting



7. Enter the dataset name and description, set extended information, and click **OK** to publish the dataset.

   If the task status is **Succeeded**, the data publishing task is successfully executed. You can choose **Data Engineering** > **Data Publishing** > **Datasets** in the navigation pane and click the **Published Dataset** tab to view the published dataset.

8. On the **Publish Task** page, click the task ID to view the task details. On the **Data Publish Detail** page, you can click the **Basic Info** and **Log Management** tabs. The **Basic Info** tab page contains the **Job Detail Info**, **Configuration Data**, **Data Source**, and **Generate Dataset** areas. In the **Generate Dataset** table, the number of records in the training set and validation set is displayed, as shown in **Figure 4-94**.

Figure 4-93 Viewing task details

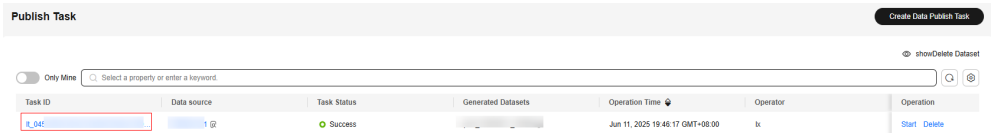
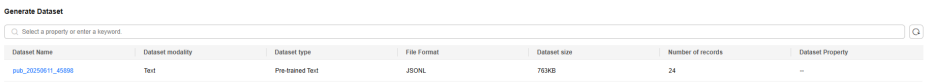


Figure 4-94 Example of the Generate Dataset table



4.6.4 Publishing Video Datasets

4.6.4.1 Evaluating Video Datasets

Before publishing a video dataset, ModelArts Studio allows you to evaluate the dataset to optimize data quality, ensure that the data meets high standards, and improve model performance.

If data evaluation is not required, skip this section and proceed with **Publishing Video Datasets**.

Creating Evaluation Standards for Video Datasets

ModelArts Studio provides preset evaluation standards for video datasets, covering multiple quality dimensions such as video definition, frame rate, integrity, and label accuracy. You can directly use those standards or create your evaluation standards.

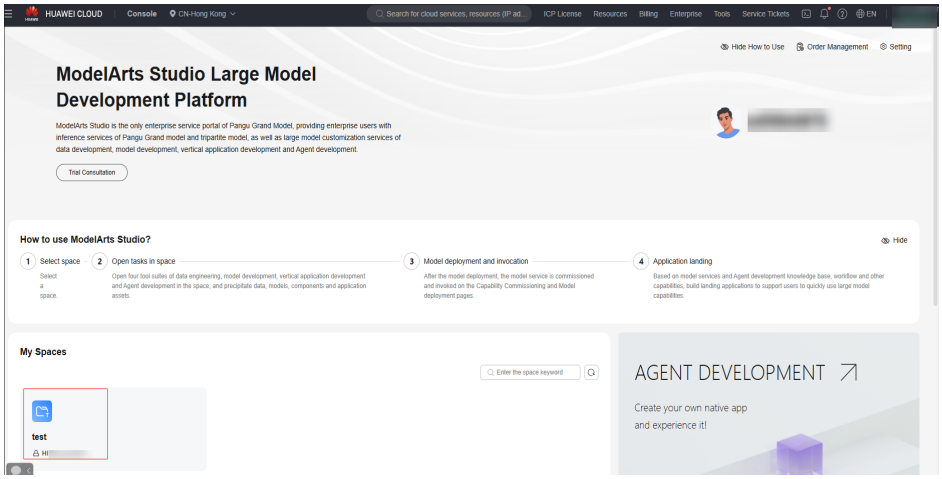
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

If you want to use the preset evaluation standards, skip this section and proceed with **Creating a Video Dataset Evaluation Task**.

To create evaluation standards for video datasets, perform the following steps:

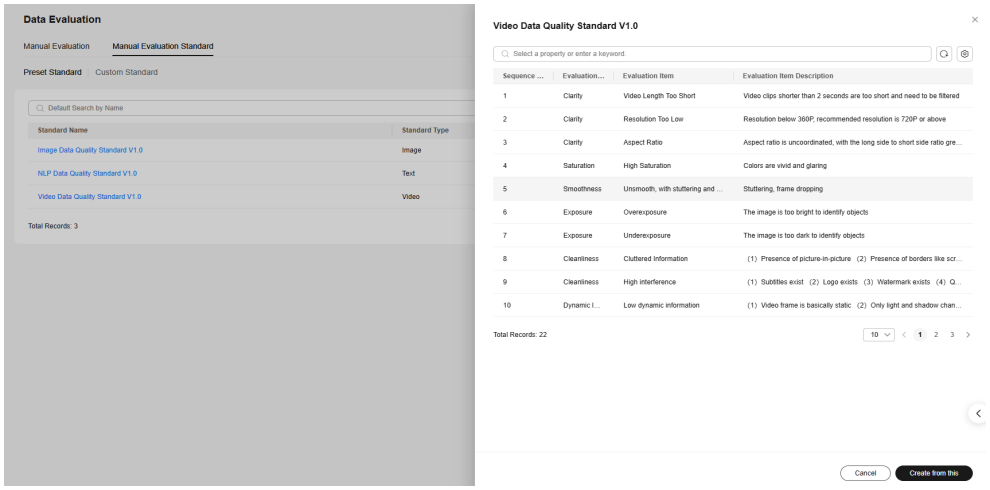
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-95 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Data Evaluation**. On the **Manual Evaluation Standard** tab page, the preset video dataset evaluation standard **Video Data Quality Standard V1.0** are displayed. You can click the evaluation standard name to view the specific evaluation items.

   Figure 4-96 Preset evaluation standards for video datasets

3. On the **Manual Evaluation Standard** page, click **Create custom standards**, select **Preset Standard**, and set the evaluation standard name and description.
4. Edit evaluation items.

   You can delete evaluation items or create custom evaluation items as required. When creating a custom evaluation item, ensure that the evaluation

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

type, evaluation item, and evaluation item description are clear and unambiguous.

5. Click **Complete Creation**.

   After evaluation standards are created, you can view, edit, and delete them on the **Manual Evaluation Standard** page.

Creating a Video Dataset Evaluation Task

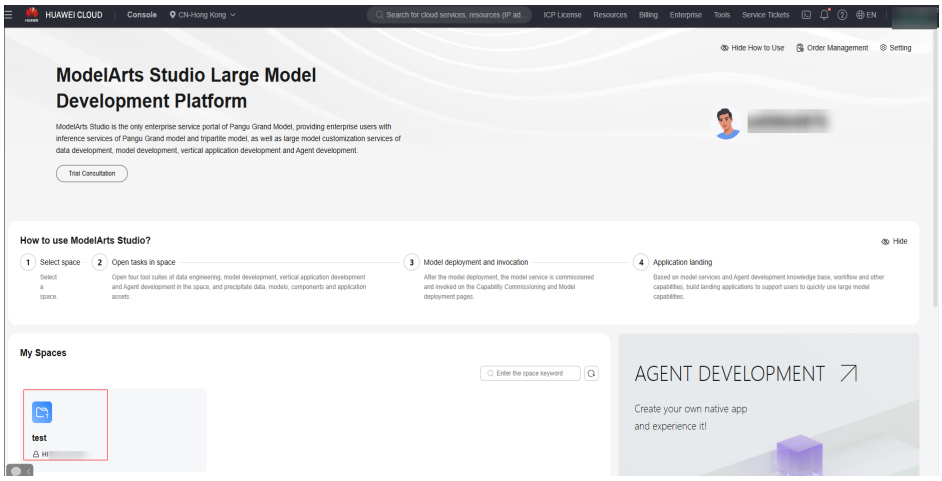
Only processed datasets can be evaluated.

Before creating a video dataset evaluation task, generate a processed dataset by referring to **Processing Video Datasets**.

To create a video dataset evaluation task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-97 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Data Evaluation**. On the displayed page, click **Create Evaluation Task** in the upper right corner.
3. Select the dataset to be evaluated and set the number of samples.
4. Click **Next Step** and select the evaluation standard. Click **Next Step** to select the evaluator, and click **Next Step** to enter the task name.
5. Click **Complete Creation** to go to the **Data Evaluation** page. After the evaluation task is created, the status becomes **Created**.
6. Click **Evaluate** in the **Operation** column.
7. On the evaluation page, label the problems of the current data by referring to the evaluation items. If the data meets the requirements, click **Passed**. If the data does not meet the requirements, click **Not passed**.
8. After all data is evaluated, check that the evaluation progress is 100% on the **Manual Evaluation** page.

   Click **Report** in the **Operation** column to view the dataset quality evaluation report.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4.6.4.2 Publishing Video Datasets

Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations.

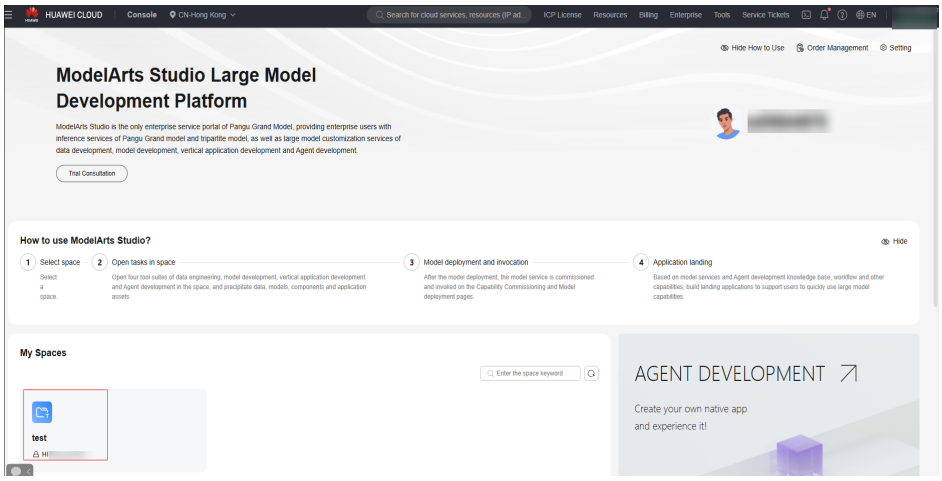
Video datasets can be published only in standard format.

Creating a Video Dataset Publishing Task

To create a video dataset publishing task, perform the following steps:

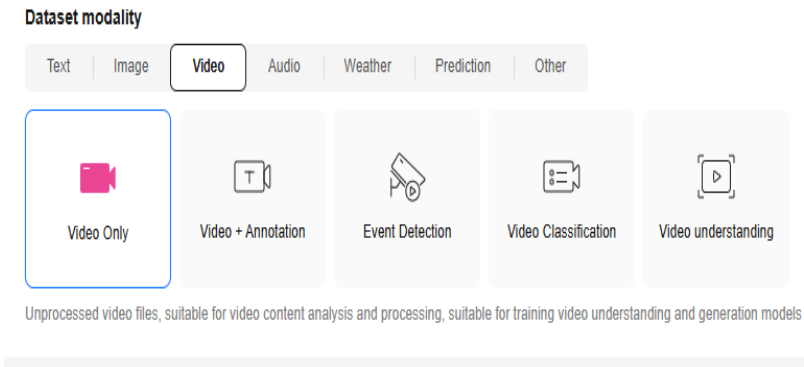
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-98 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
3. On the **Create Data Publish Task** page, select a dataset modality, for example, **Video** > **Video Only**.

   Figure 4-99 Selecting dataset modality

4. Select a dataset and click **Next**.
5. In the **Basic Configuration** area, select the data usage, dataset visibility, and application scenario. Currently, video datasets can be released only in standard format.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

6. Enter the dataset name and description, set extended information, and click **OK** to publish the dataset.

   If the task status is **Succeeded**, the data publishing task is successfully executed. You can choose **Data Engineering** > **Data Publishing** > **Datasets** in the navigation pane and click the **Published Dataset** tab to view the published dataset.

4.6.5 Publishing Audio Datasets

Data publishing refers to publishing a dataset in a specific format as a published dataset for subsequent model training operations.

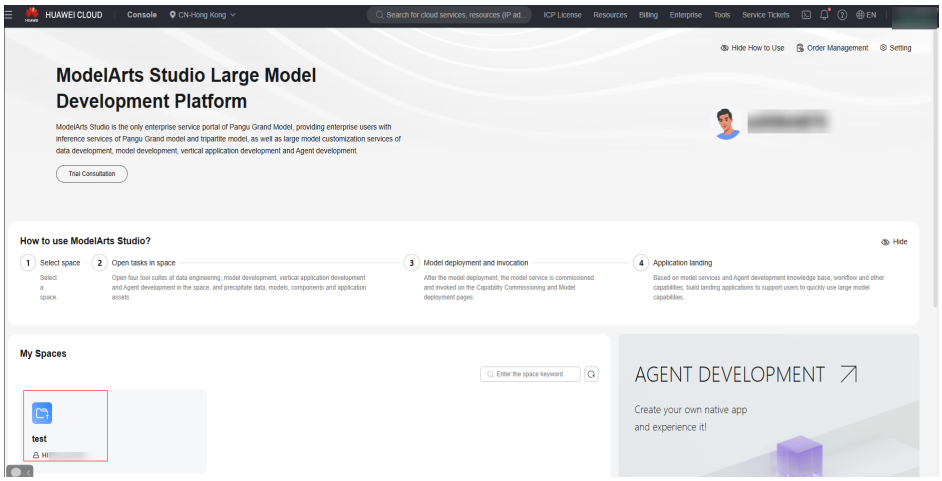
Audio datasets can only be published in standard format.

Creating an Audio Dataset Publishing Task

To create a video dataset publishing task, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

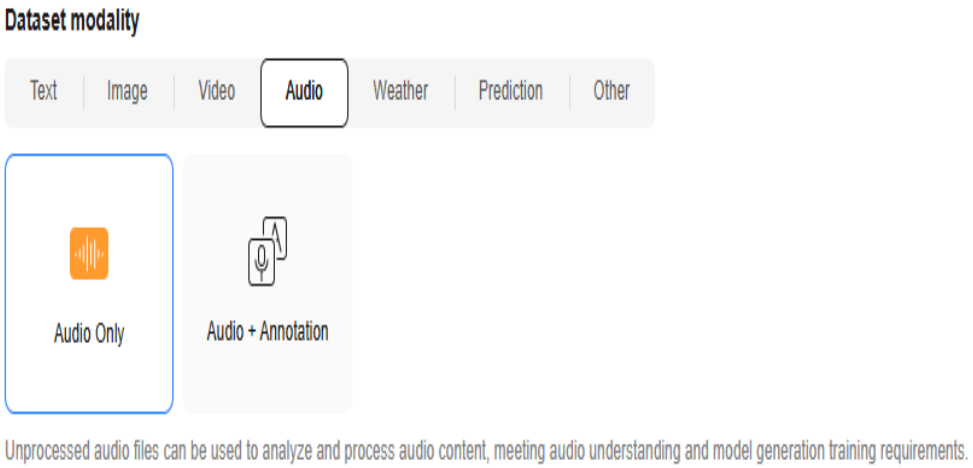
   Figure 4-100 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
3. On the **Create Data Publish Task** page, select a dataset modality, for example, **Audio** > **Audio Only**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-101 Selecting dataset modality



4. Select a dataset and click **Next**.
5. In the **Basic Configuration** area, select the data usage, dataset visibility, and application scenario. Currently, audio datasets can be released only in standard format.
6. Enter the dataset name and description, set extended information, and click **OK** to publish the dataset.

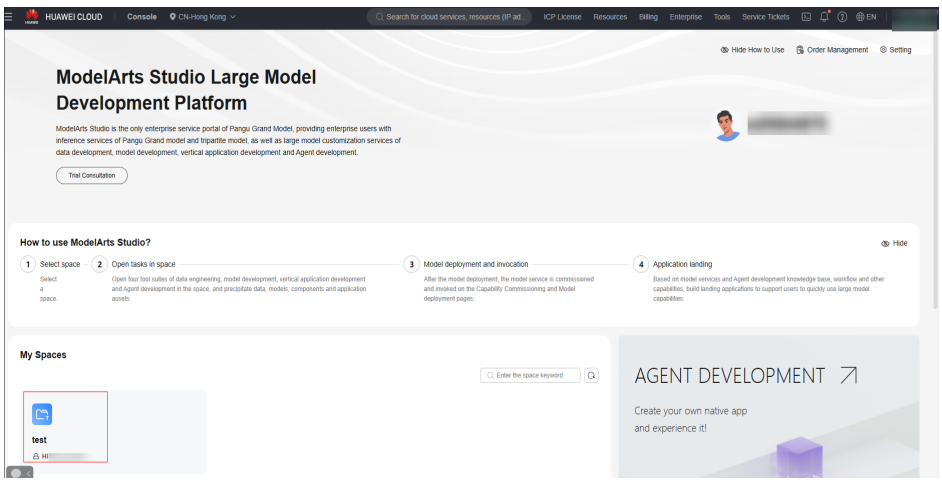
   If the task status is **Succeeded**, the data publishing task is successfully executed. You can choose **Data Engineering** > **Data Publishing** > **Datasets** in the navigation pane and click the **Published Dataset** tab to view the published dataset.

4.6.6 Publishing Weather Datasets

To publish a weather dataset, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

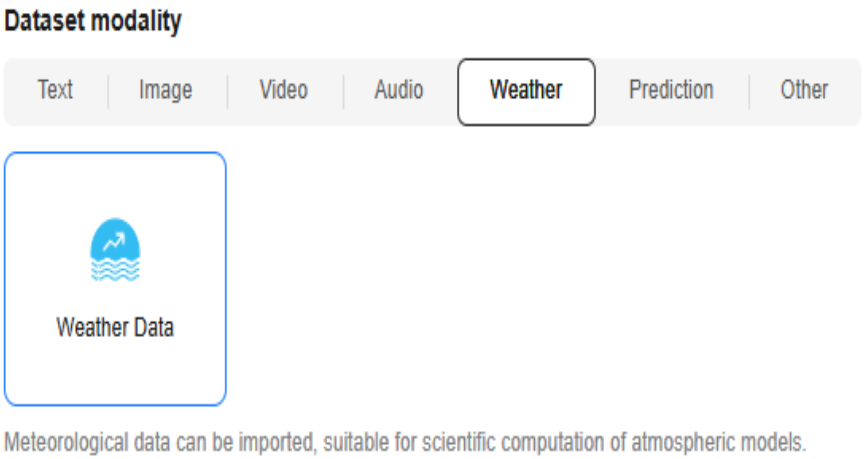
Figure 4-102 My Spaces



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
3. On the **Create Data Publish Task** page, select a dataset modality, for example, **Weather** > **Weather Data**.

   Figure 4-103 Selecting dataset modality

4. Select a dataset and click **Next**.
5. In the **Basic Configuration** area, select the data usage and dataset visibility.
6. Enter the dataset name and description, set extended information, and click **OK** to publish the dataset.

   If the task status is **Succeeded**, the data publishing task is successfully executed. You can choose **Data Engineering** > **Data Publishing** > **Datasets** in the navigation pane and click the **Published Dataset** tab to view the published dataset.

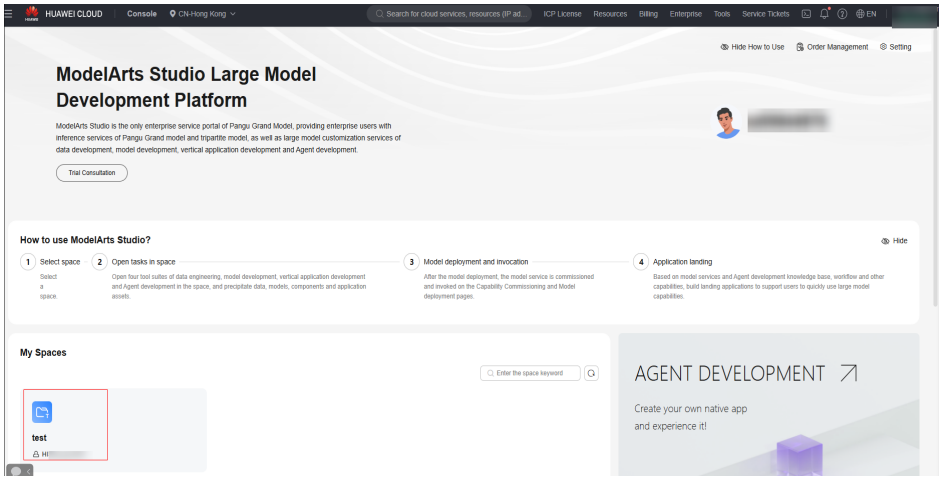
4.6.7 Publishing Other Datasets

To publish a dataset of other types, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

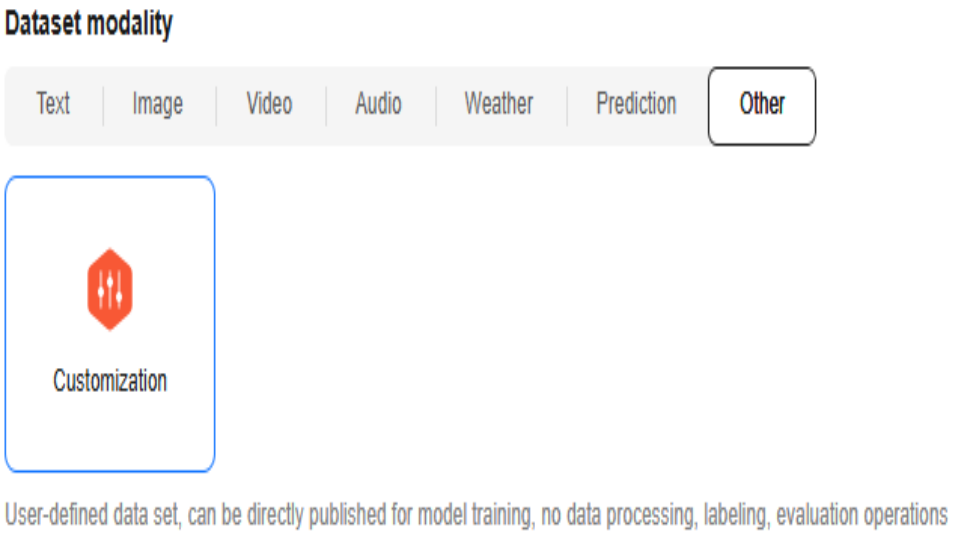
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-104 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
3. On the **Create Data Publish Task** page, select a dataset modality, for example, **Other** > **Customization**.

Figure 4-105 Selecting dataset modality



4. Select a dataset and click **Next**.
5. In the **Basic Configuration** area, select the data usage and dataset visibility.
6. Enter the dataset name and description, set extended information, and click **OK** to publish the dataset.

   If the task status is **Succeeded**, the data publishing task is successfully executed. You can choose **Data Engineering** > **Data Publishing** > **Datasets** in the navigation pane and click the **Published Dataset** tab to view the published dataset.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

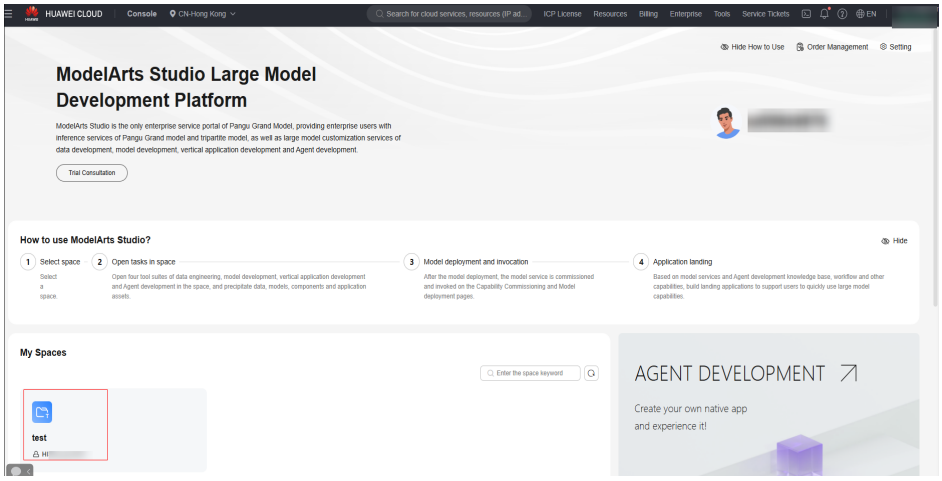
4.6.8 Managing Published Datasets

If you have completed data combination or data circulation tasks for a dataset, click **Generate** in the corresponding task list to generate a published dataset. The generated datasets are managed by the platform in a unified manner and used for subsequent publishing tasks.

The platform allows you to view basic information and data lineage of published datasets. The procedure is as follows:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 4-106 My Spaces

2. In the navigation pane, choose **Data Engineering** > **Data Management** > **Datasets**. Click the **Published Dataset** tab.
3. Click the dataset name to view the basic information, data preview, data lineage, and operation records of the published dataset.
   - On the **Basic Information** tab page, you can view the data details, data source, and extended information. In the **Extended Info** area, you can set dataset properties as required, including the dataset property name, industry, language, and custom tag.
   - On the **Data Preview** tab page, you can view the published data.
   - On the **Data Lineage** tab page, view the operations performed on the dataset, such as import and synthesis.
   - On the **Operation Record** tab page, view the operations and status of the dataset.
4. Click **Delete** in the **Operation** column to delete unnecessary datasets.
   - To restore a deleted dataset, click **Show Deleted Data** in the upper right corner. In the list of deleted datasets, restore the dataset.
   - To permanently delete a dataset, click the dataset name to go to the dataset details page, confirm the dataset content, and delete it permanently.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

NOTE

Deleting a published dataset is a high-risk operation. Before deleting a dataset, ensure that it is no longer used.

4.7 Converting the Dataset Format

4.7.1 Constraints

Supported Dataset Types

Custom dataset structure: Image suffixes (jpg, jpeg, png, bmp, tif, and tiff) are supported.

Assume that the name of the customized dataset file is test_customized dataset.

+---test_Custom dataset

| +---Image 001 # Python operator directory

| +---Image 002 # Python operator directory

| +---Image 003 # Python operator directory

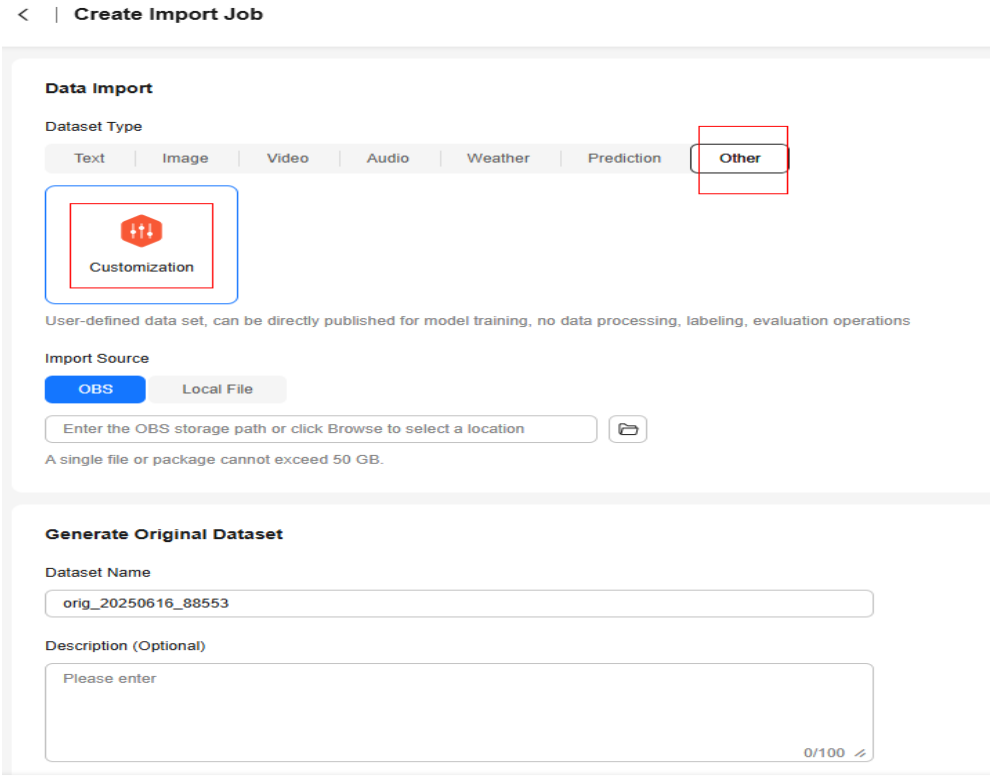
| +--- CSV file

Supported CV Dataset Types

Only custom datasets can be converted into image (object detection and image classification) datasets.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

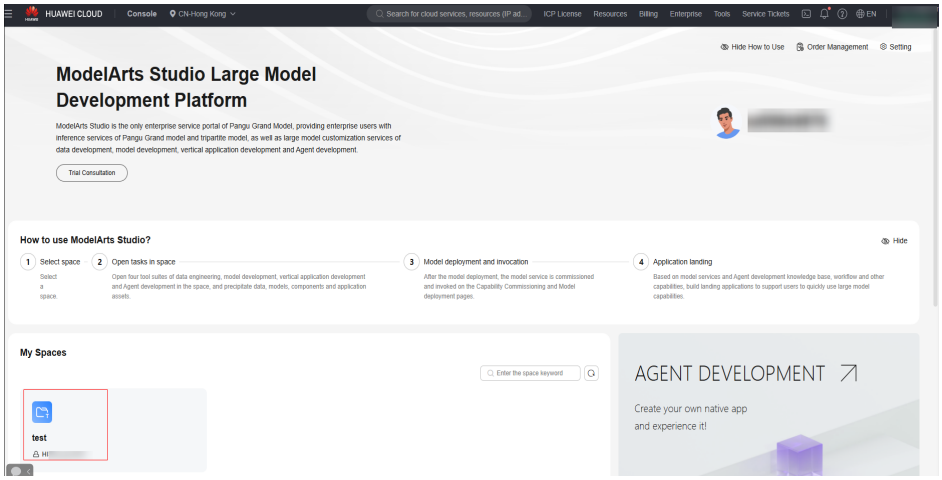
Figure 4-107 Importing a Custom Dataset



4.7.2 Format Conversion Process

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

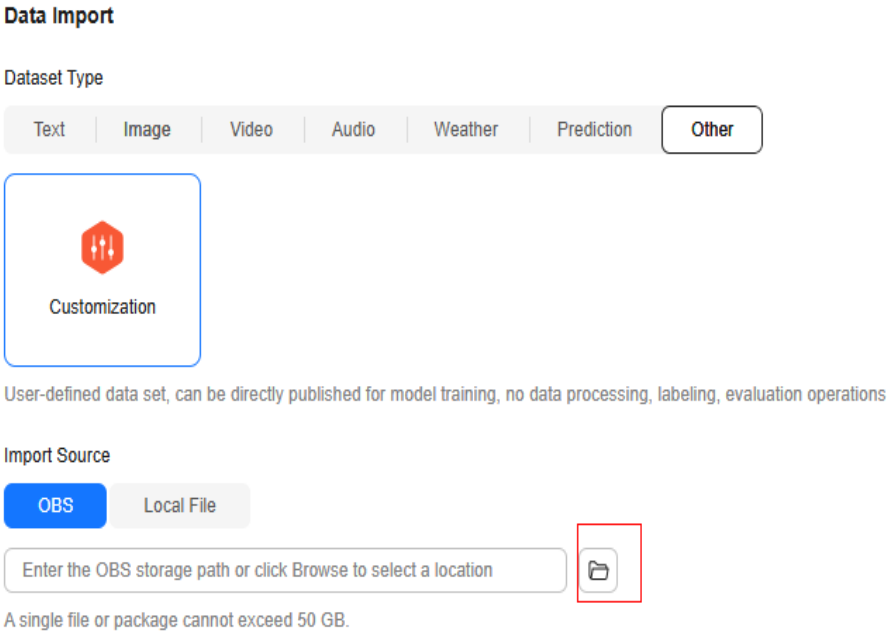
Figure 4-108 My Spaces



2. In the navigation pane, choose **Data Engineering** > **Data Acquisition** > **Data Import**. On the displayed page, click **Create Import Job** in the upper right corner.
3. On the Create Import Job page, set **Dataset Type** to **Other** > **Customization**. Import the dataset to be converted, as shown in **Figure 4-109**.

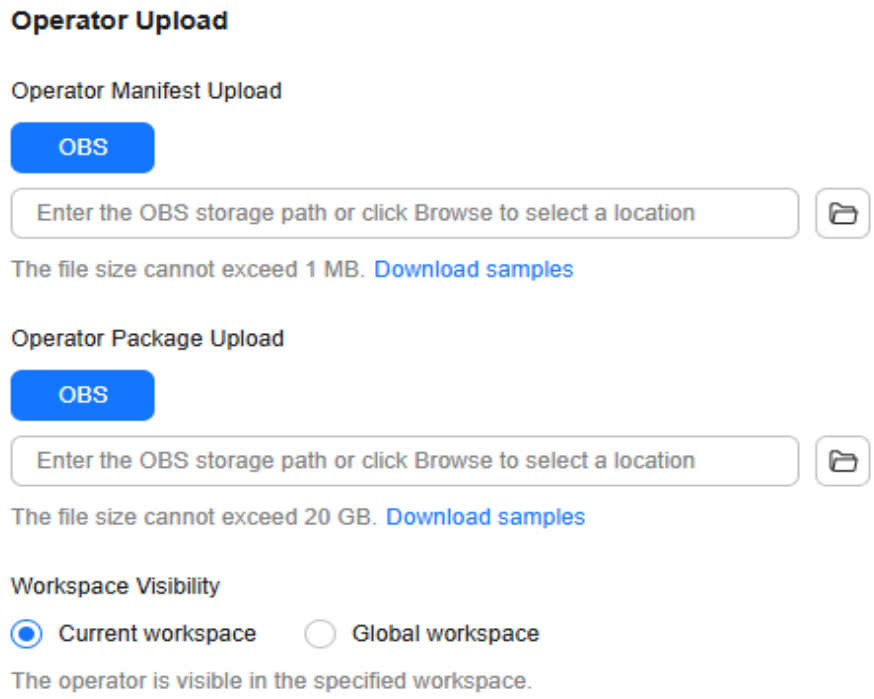
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 4-109 Importing data



4. In the navigation pane, choose **Data Engineering** > **Data Processing**. On the **Processing Tasks** page, click **Manage Processing Operator**, and click the **Custom** tab page, as shown in **Figure 4-110**.

Figure 4-110 Creating a custom operator



5. Import the corresponding YML file and TAR package to build a format conversion operator.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

6. In the navigation pane, choose **Data Engineering** > **Data Processing**. On the **Processing Tasks** page, click **Create Processing Job**.
7. Select a dataset and a conversion operator to start the format conversion task.

NOTE

1. Object detection and image classification support the JPG, JPEG, BMP, PNG, TIF, and TIFF formats.
2. The CSV file is in UTF-8 format and the content to be read is on the first sheet.

4.7.3 Conversion Operators

The conversion operator construction requires a YML configuration file and the Python script TAR package for implementing functions.

1. YML file example:

```
### Python operator example

id: FormatConvert2
name: Custom operator format conversion
description: Format conversion using custom operators
author: Huawei Cloud Technology Co., Ltd.
tags:
  language:
    - zh
    - en
  format:
    - CSV
    - TIF
  category: Data conversion
  modal:
    - IMAGE
runtime:
  cpu-arch:
    - X86
  resources:
    - cpu: 1
      memory: 4096
  environment: python
  entrypoint: process.py
  auto-data-loading: false # Whether to automatically load data
arguments:
  - key: operateType
  name: Operation type
  type: ENUM  #Single choice
    items:
      - name: Object detection
        value: object_detection
      - name: Image classification
        value: image_classification
    required: true
    visible: true
    default: object_detection
    """
    return "parquet"
```

2. TAR package:

Assume that the operator package name is **operator-package.tar**. The directory structure after the operator package is decompressed is as follows:

+---operator-package

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

| +--- program_package # Python operator directory

| | +---dependency

| | | +---requirements.txt # (Optional) Requirement description

| | +--- install.sh # (Optional) Installation script

| | +--- process.py # (Mandatory) Operator code

process.py file for format conversion

```
import PIL.Image
import PIL.ImageOps
import numpy as np
import ast
import os
import pandas as pd
import xml.etree.ElementTree as ET
from xml.dom import minidom
import PIL
import moxing as mox
import ma_utils as utils

logger = utils.FileLogger.get_logger()

import argparse


arg_parser = argparse.ArgumentParser()
arg_parser.add_argument("--operateType", type=str, required=True, help="")
arg_parser.add_argument("--input_obs_path", type=str, required=True, help="")
arg_parser.add_argument("--output_obs_path", type=str, required=True, help="")


def prettify_xml(elem):
    rough_string = ET.tostring(elem, 'utf-8')
    reparsed = minidom.parseString(rough_string)
    return reparsed.toprettyxml(indent="  ")


def create_pascal_voc_xml(folder, filename, path, width, height, depth, objects, deu, db, dp, ls, lv, lse, lc, srct,
                          srrt, src, srr, rct, rrt, rc, rr, dest__engine_uuid=None):
    annotation = ET.Element('annotation')
    # UUID of the destination engine
    dest__engine_uuid = ET.SubElement(annotation, 'DEST_ENGINE_UUID')
    dest__engine_uuid.text = deu
    # Destination bucket (logical partition)
    dest__engine_uuid = ET.SubElement(annotation, 'DEST_BUCKET')
    dest__engine_uuid.text = db
    # Destination path
    dest_path = ET.SubElement(annotation, 'DEST_PATH')
    dest_path.text = folder
    # File name
    filename_element = ET.SubElement(annotation, 'FILE_NAME')
    filename_element.text = filename
    # Full path of the label group
    dest_path = ET.SubElement(annotation, 'DEST_PATH')
    dest_path.text = dp
    # Full path of the label group
    label_value = ET.SubElement(annotation, 'LABEL_VALUE')
    label_value.text = lv
    # Label shape point/line/rectangle (two points) /rectangle (bbox) /rectangle (center point) /polygon/
contour
    label_shape = ET.SubElement(annotation, 'LABEL_SHAPE')
    label_shape.text = ls
    # Label shape point/line/rectangle (two points) /rectangle (bbox) /rectangle (center point) /polygon/
contour
    label_coordinate = ET.SubElement(annotation, 'LABEL_COORDINATE')
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
    label_coordinate.text = lc
    # Time when a source record is created
    rec_time = ET.SubElement(annotation, 'SOURCE_REC_CREATE_TIME')
    rec_time.text = srct
    # Time when a source record is modified
    rec_rev_time = ET.SubElement(annotation, 'SOURCE_REC_REVISE_TIME')
    rec_rev_time.text = srrt
    # Creator of the source record
    rec_ctor = ET.SubElement(annotation, 'SOURCE_REC_CREATOR')
    rec_ctor.text = src
    # Modifier of the source record
    rec_r_ctor = ET.SubElement(annotation, 'SOURCE_REC_REVISOR')
    rec_r_ctor.text = srr
    # Time when a source record is created
    d_cd_time = ET.SubElement(annotation, 'REC_CREATE_TIME')
    d_cd_time.text = rct
    # Time when a record is modified
    r_time = ET.SubElement(annotation, 'REC_REVISE_TIME')
    r_time.text = rrt
    # Record creator
    rec_c = ET.SubElement(annotation, 'REC_CREATOR')
    rec_c.text = rc
    # Record modifier
    recd_time = ET.SubElement(annotation, 'REC_REVISOR')
    recd_time.text = rr
    tree = ET.ElementTree(annotation)
    return tree
def create_pascal_voc_xml1(folder, filename, path, width, height, depth, objects):
    annotation = ET.Element('annotation')

    folder_element = ET.SubElement(annotation, 'folder')
    folder_element.text = folder

    filename_element = ET.SubElement(annotation, 'filename')
    filename_element.text = filename

    path_element = ET.SubElement(annotation, 'path')
    path_element.text = path

    source = ET.SubElement(annotation, 'source')
    database = ET.SubElement(source, 'database')
    database.text = 'Unknown'

    size = ET.SubElement(annotation, 'size')
    width_element = ET.SubElement(size, 'width')
    width_element.text = str(width)
    height_element = ET.SubElement(size, 'height')
    height_element.text = str(height)
    depth_element = ET.SubElement(size, 'depth')
    depth_element.text = str(depth)

    segmented = ET.SubElement(annotation, 'segmented')
    segmented.text = '0'

    for obj in objects:
        obj_element = ET.SubElement(annotation, 'object')
        name_element = ET.SubElement(obj_element, 'name')
        name_element.text = obj['name']

        pose = ET.SubElement(obj_element, 'pose')
        pose.text = 'Unspecified'

        truncated = ET.SubElement(obj_element, 'truncated')
        truncated.text = '0'

        difficult = ET.SubElement(obj_element, 'difficult')
        difficult.text = '0'

        bndbox = ET.SubElement(obj_element, 'bndbox')
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
        xmin = ET.SubElement(bndbox, 'xmin')
        xmin.text = str(obj['xmin'])
        ymin = ET.SubElement(bndbox, 'ymin')
        ymin.text = str(obj['ymin'])
        xmax = ET.SubElement(bndbox, 'xmax')
        xmax.text = str(obj['xmax'])
        ymax = ET.SubElement(bndbox, 'ymax')
        ymax.text = str(obj['ymax'])

    tree = ET.ElementTree(annotation)
    return tree

def get_value(row, name):
    value = '' if pd.isna(row[name]) else str(row[name])
    return value


def copy_local_to_obs(local_file_path, obs_file_path):
    mox.file.copy(local_file_path, obs_file_path)


def csv_to_pascal_voc(args, csv_file, local_tif_file_path_dict, local_output_file_dir):
    operate_type = getattr(args, 'operateType', 'object_detection')
    logger.info(operate_type)
    df = pd.read_csv(csv_file)
    grouped = df.groupby('FILE_NAME')

    for filename, group in grouped:
        if local_tif_file_path_dict.get(filename, None):
            if operate_type == 'object_detection':

                objects = []
                for _, row in group.iterrows():
                    bndbox = ast.literal_eval(row['LABEL_COORDINATE'])
                    label = row['LABEL_VALUE']
                    obj = {
                        'name': label,
                        'xmin': bndbox[0],
                        'ymin': bndbox[1],
                        'xmax': bndbox[2],
                        'ymax': bndbox[3]
                    }
                    objects.append(obj)

                width, height, depth = 0, 0, 0

                folder = os.path.join(row['DEST_BUCKET'], row['DEST_PATH'])
                path = os.path.join(folder, filename)
#                 real_path = os.path.join(test_root, path)
                with PIL.Image.open(local_tif_file_path_dict.get(filename)) as img:
                    width, height = img.size
                    depth = len(img.getbands())
                tree = create_pascal_voc_xml1(folder, filename, path, width, height, depth, objects)
                pretty_xml = prettify_xml(tree.getroot())

                local_output_xml_file_path = os.path.join(local_output_file_dir, os.path.splitext(filename)[0] +
'.xml')
                with open(local_output_xml_file_path, 'w', encoding='utf-8') as f:
                    f.write(pretty_xml)
                fname = replacetype(filename,'1')
                copy_local_to_obs(local_output_xml_file_path,os.path.join(args.output_obs_path, fname))
                copy_local_to_obs(local_tif_file_path_dict.get(filename), os.path.join(args.output_obs_path,
filename))
            elif operate_type == 'image_classification':
                for _, row in group.iterrows():
                    logger.info(f'-------------Cycle row--')

                lv = get_value(row, 'LABEL_VALUE')
                txtstr = create_pascal_voc_txt(lv)
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
            fname = replacetype(filename,'2')
            local_output_txt_file_path = os.path.join(local_output_file_dir, fname)
            with open(local_output_txt_file_path, 'w', encoding='utf-8') as f:
                f.write(txtstr)
            copy_local_to_obs(local_output_txt_file_path,
                              os.path.join(args.output_obs_path, fname))
            copy_local_to_obs(local_tif_file_path_dict.get(filename), os.path.join(args.output_obs_path,
filename))




def replacetype(filename,type):
    if type == '1':
        filename = filename.replace('.jpg', '.xml')
        filename = filename.replace('.bmp', '.xml')
        filename = filename.replace('.jpeg', '.xml')
        filename = filename.replace('.png', '.xml')
        filename = filename.replace('.JPG', '.xml')
        filename = filename.replace('.BMP', '.xml')
        filename = filename.replace('.JPEG', '.xml')
        filename = filename.replace('.PNG', '.xml')
    elif type == '2':
        filename = filename.replace('.jpg', '.txt')
        filename = filename.replace('.bmp', '.txt')
        filename = filename.replace('.jpeg', '.txt')
        filename = filename.replace('.png', '.txt')
        filename = filename.replace('.JPG', '.txt')
        filename = filename.replace('.BMP', '.txt')
        filename = filename.replace('.JPEG', '.txt')
        filename = filename.replace('.PNG', '.txt')
    return filename


def create_pascal_voc_txt(lv):
    # tree='DEST_ENGINE_UUID:'+deu+'\n'+'DEST_BUCKET:'+db
#     txttree = 'DEST_ENGINE_UUID:' + deu + '\n' + 'DEST_BUCKET:' + db + '\n' + 'DEST_PATH:' + dp + '\n' +
'LABEL_SET:' + ls + '\n' + 'LABEL_VALUE:' + lv + '\n' + 'LABEL_SHAPE:' + lse + '\n' + 'LABEL_COORDINATE:' +
lc + '\n' + 'S#OURCE_REC_CREATE_TIME:' + srct + '\n' + 'SOURCE_REC_REVISE_TIME:' + srrt + '\n' +
'SOURCE_REC_CREATOR:' + src + '\n' + 'SOURCE_REC_REVISOR:' + srr + '\n' + 'REC_CREATE_TIME:' + rct +
'\n' + 'REC_REVISE_TIME:' + rrt + '\n' + 'RE#C_CREATOR:' + rc + '\n' + 'REC_REVISOR:' + rr
    txttree = lv

    return txttree


class Process:
    def __init__(self, args):
        # Create a local directory to store files transferred from OBS.
        self.local_input_file_dir = r'/tmp/obs_input/'
        os.makedirs(self.local_input_file_dir, exist_ok=True)
        # Create a local directory to store the processed files and transfer them to OBS.
        self.local_output_file_dir = r'/tmp/obs_output/'
        os.makedirs(self.local_output_file_dir, exist_ok=True)
        # Set parameters. Set default values for operator parameters.
        self.args = args
        self.input_obs_path=args.obs_input_path
        self.output_obs_path=args.obs_output_path
        self.args.input_obs_path=args.obs_input_path
        self.args.output_obs_path=args.obs_output_path
        logger.info(f'-------Operator execution--------')

    def __call__(self, input):
        # Obtain the file name in the OBS path.
        file_name_list = mox.file.list_directory(self.args.input_obs_path, recursive=False)
        file_path_dict = {file_name: os.path.join(self.args.input_obs_path, file_name) for file_name in
file_name_list}
        # Process each file.
        local_tif_file_path_dict, local_csv_file_path = {}, ''
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
    for file_name, obs_file_path in file_path_dict.items():
    # Copy the file from OBS to the local host.

#         if file_name.endswith('.jpg'):
#             local_tif_file_path = os.path.join(self.local_input_file_dir, file_name)
#             logger.info(f'---------------4'+local_tif_file_path)
#             local_tif_file_path_dict[file_name] = local_tif_file_path
#             mox.file.copy(obs_file_path, local_tif_file_path)
#         else:
#             local_csv_file_path = os.path.join(self.local_input_file_dir, file_name)
#             logger.info(f'---------------5'+local_csv_file_path)
#             mox.file.copy(obs_file_path, local_csv_file_path)

        if file_name.endswith('.csv'):

            local_csv_file_path = os.path.join(self.local_input_file_dir, file_name)
            mox.file.copy(obs_file_path, local_csv_file_path)
        else:
            local_tif_file_path = os.path.join(self.local_input_file_dir, file_name)
            local_tif_file_path_dict[file_name] = local_tif_file_path
            mox.file.copy(obs_file_path, local_tif_file_path)

    csv_to_pascal_voc(self.args, local_csv_file_path, local_tif_file_path_dict, self.local_output_file_dir)
```

4.8 Common Errors and Solutions for Data Engineering

Table 4-45 lists the common errors and solutions for data engineering.

Table 4-45 Common errors and solutions for data engineering

Function	Error	Solution
Data Acquisition	File format mismatch, require [{0}].	Check whether the data used for creating the dataset is in the format required by the platform.
	Verification failed. Please check the content format is consistent with the template requirements.	Check whether the data used for creating the dataset is in the format required by the platform.
	content type [%s] not support, only [%s] support.	The content in the dataset is not supported. Ensure that the format of the uploaded data is the same as that required by the platform.
	get obs bucket folders error.	Check whether the OBS service is normal and whether data in the OBS bucket can be accessed.
Data Processing	dataset is not online.	The dataset used for data processing has not been brought online. Bring the dataset online first.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Function	Error	Solution
	invalid obs path.	Check whether the OBS path corresponding to the dataset is valid and can be accessed.
Data Labeling	annotate data not exist.	Check whether the labeled dataset exists and has been deleted.
	obs url invalid.	Check whether the OBS path corresponding to the dataset is valid and can be accessed.
	data management query dataset data invalid.	Check whether the labeled dataset exists and has been deleted.
	dataset obs file empty.	Check whether the dataset file exists in the original OBS bucket.
	download obs file failed.	Check whether the network is normal and whether data in the OBS bucket can be accessed.
Data Evaluation	annotate type is invalid.	Check whether the data labeling type and data labeling requirements in the uploaded data are consistent with those required by the platform.
	annotate data not exist.	The evaluation dataset does not exist. Check whether the dataset has been successfully imported and whether the OBS bucket is empty.
	obs url invalid.	Check whether the OBS path corresponding to the dataset is valid and can be accessed.
	standard item not exist.	Check whether the evaluation criteria exist and have been deleted.
	the corresponding data has been marked as deleted.	Check whether the evaluation dataset is deleted.
	the current data not exist.	Check whether the evaluation dataset exists and has been deleted.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Function	Error	Solution
	dataset file type does not match standard file type.	Check whether the type of the uploaded dataset file is the same as the standard file type required by the platform.
	data management query dataset data invalid.	Check whether the dataset contains data in incorrect formats.
	dataset obs file empty.	Check whether the dataset file exists in the original OBS bucket.
Data Publishing	Dataset [%s] is not found.	Check whether the dataset exists.
	Dataset [%s] status is invalid.	Check whether the dataset status is normal.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

5 Developing a DeepSeek Model

5.1 DeepSeek Models

The platform provides the DeepSeek-R1-32K and DeepSeek-V3-32K models.

DeepSeek-R1

DeepSeek-R1 671B, a full-scale version of the DeepSeek-R1 model, delivers on-par performance with OpenAI's o1 in a range of tasks, such as mathematics, coding, and natural language processing. DeepSeek-R1-32K-0.0.1 supports inference deployment.

The platform provides preconfigured inference services. You can call the inference services through the experience center or APIs.

The optimal model has been pre-built. There is no need to compress the model before deployment.

Table 5-1 Inference features

Inference Precision	INT8
Sequence length	32K
Minimal inference unit	16
Concurrency	256

DeepSeek-V3

DeepSeek-V3 is a strong Mixture-of-Experts (MoE) model with 671B total parameters (with 37B activated for each token). It demonstrates remarkable performance across a wide range of tasks, such as long-context processing, code generation, mathematical reasoning, general knowledge tasks, and Chinese language understanding. DeepSeek-V3-32K-0.0.1 supports inference deployment.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

The optimal model has been pre-built. There is no need to compress the model before deployment.

Table 5-2 Inference features

Inference Precision	INT8
Sequence length	32K
Minimal inference unit	16
Concurrency	256

5.2 Using Data Engineering to Build a DeepSeek Model Dataset

Process of Building a DeepSeek Model Dataset

Table 5-3 describes how to use data engineering to build a third-party model dataset on ModelArts Studio.

Table 5-3 Process of building a third-party model dataset

Procedure	Step	Description	Reference
Importing data to the Pangu platform	Creating an import task	Import data stored in OBS or local data into the platform for centralized management, facilitating subsequent processing or publishing. **NOTE** When importing a dataset, set the dataset type to **Single Round QA**.	**Importing Data to the Pangu Platform**
Processing other datasets	Processing other datasets	Use custom processing operators to preprocess data, ensuring it meets the model training standards and service requirements.	**Processing Other Datasets**
Publishing other datasets	Publishing other datasets	Data publishing refers to publishing a single dataset in a specific format as a published dataset for subsequent model training operations.	**Publishing Other Datasets**

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Requirements on DeepSeek Datasets

Table 5-4 Dataset format for DeepSeek models

Model Type	Training Type	Data Volume	Dataset Format	Description
DeepSeek-R1-32K DeepSeek-V3-32k	Pre-training	> 15B tokens	JSONL	JSONL format: **text** indicates the text data used for pre-training. The following is an example: {"text":"Pangu models include the NLP model, multimodal model, CV model, scientific computing model, and prediction model."}
DeepSeek-V3-32k	Fine-tuning (single-turn Q&A)	10,000 to 1,000,000 data records	JSONL	JSONL format: The data consists of Q&A pairs. **context** and **target** indicate the question and answer, respectively. The following is an example: {"context":"Hello, please introduce yourself.","target":"I am a Pangu model."}
DeepSeek-R1-32K	Fine-tuning (single-turn Q&A)	10,000 to 1,000,000 data records	JSONL	JSONL format: The data consists of Q&A pairs. **context** and **target** indicate the question and answer, respectively. In addition to the model answer, **target** should include <think>\nReasoning process\n</think>, which indicates the reasoning process of the model. {"context":"Question","target":"<think>\nReasoning process\n</think>\n\nModel answer"} Example: {"context":"What is 2 + 2?","target":"<think>\nAccording to the basic arithmetic definition, the sum of 2 and 2 is the combination of two values.\nVerification is performed through object counting (e.g., two apples plus two apples) or by shifting values on a number line (e.g., from 0 to 2 to 4).\nThis value adheres to the natural number axiomatic system. For example, the successor of 2 is 3, and the successor of 3 is 4.\n</think>\n\n**Answer:** 2 + 2 = 4"}

5.3 Training a DeepSeek Model

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

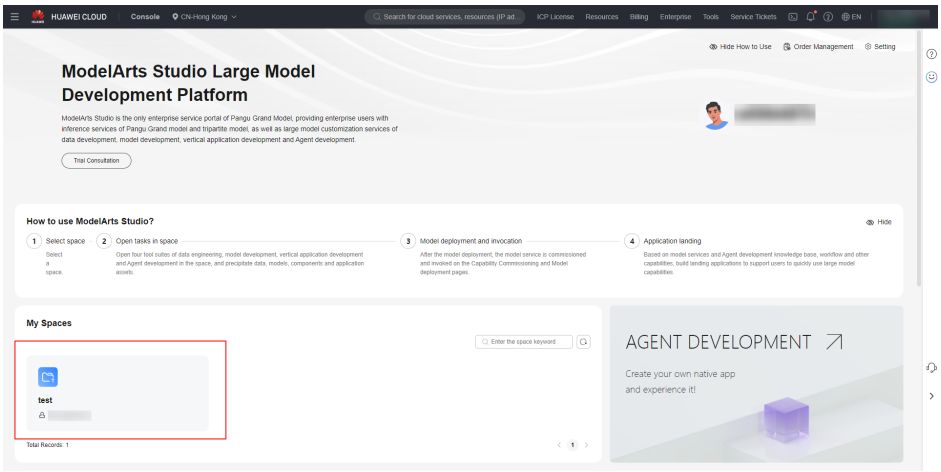
5.3.1 Creating a DeepSeek Model Training Job

Pre-training

To create a DeepSeek model pre-training job, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 5-1 My Spaces



2. In the navigation pane, choose **Model Development** > **Model Training**. Click **Create Training Job** in the upper right corner.
3. On the **Create Training Job** page, set training parameters by referring to **Table 5-5**.

Table 5-5 Parameters for pre-training the DeepSeek model

Category	Training Parameter	Description
Training configuration	Select Model	You can modify the following information: • **Sources**: Select **Model Square**. • **Type**: Select **NLP** and select the base model and version (DeepSeek-V3-32K or DeepSeek-R1-32K) used for training.
	Type	Select **Pre-training**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	Advanced Settings > checkpoints	checkpoints: During a model training job, checkpoints are used to store the model weight and status. • **Close**: After this function is disabled, checkpoints are not saved and training cannot be resumed based on checkpoints. • **Automatic**: All checkpoints during training are automatically saved. • **Custom**: A specified number of checkpoints are saved based on the settings.
Training Parameters	epochs	Dataset iterations during model training.
	learning_rate	It determines the step size at each training iteration. If the learning rate is too high, the model never converges. If the learning rate is too low, the model can take a long time to converge.
	batch_size	It controls the number of samples used in one iteration of model training. A larger batch size leads to more stable gradients but consumes more GPU memory. This may cause out-of-memory (OOM) errors due to hardware limitations and extend the training time.
	sequence_length	It specifies the maximum length of a single training data record. Data that exceeds the length will be truncated during training.
	warmup	It controls the duration of the warm-up phase relative to the entire training process. During this phase, the learning rate starts low, allowing the model to stabilize gradually. Once stabilized, the model is then trained at a preset higher learning rate, which accelerates convergence and enhances performance.
	lr_decay_ratio	It controls the learning rate to ensure the model converges more stably. The minimum learning rate will not be lower than the product of the **Learning Rate** and this parameter. If the learning rate does not decay, set this parameter to **1**.
	weight_decay	It is a regularization method that reduces the size of model parameters to prevent model overfitting and improve the model generalization capability.
	checkpoint_save_strategy	It can be set to **save_checkpoint_steps** or **save_checkpoint_epoch**. It specifies whether to save checkpoint files based on the number of iteration steps or based on the number of epochs during training.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	save_checkpoint_steps or save_checkpoint_epoch	**save_checkpoint_steps** specifies the number of training steps at which the model checkpoint file is saved during training. **save_checkpoint_epoch** specifies the number of training epochs during which the model checkpoint file is saved.
	checkpoints	During a model training job, checkpoints are used to store the model weight and status. ● **Close**: After this function is disabled, checkpoints are not saved and training cannot be resumed based on checkpoints. ● **Automatic**: All checkpoints during training are automatically saved. ● **Custom**: A specified number of checkpoints are saved based on the settings.
Training data configuration	Training set	Select the training dataset.
Resource Disposition	Billing model	Billing mode of the current training job.
	Training Unit	Select the number of training units required for training the model. The minimum number of training units required for the current training is displayed.
	Single-instance training units	Select the number of training units for a single instance.
	Number of instances	Select the number of instances.
Subscription reminder	Subscription reminder	After this function is enabled, the system sends SMS or email notifications to users when the task status is updated.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
publish model	Enable automatic publishing	After this function is enabled, the final product generated after model training is automatically published as a workspace asset so that the model can be compressed, deployed, evaluated, and shared with other workspaces.
Basic Information	Name	Name of a training job.
	Description	Description of the training job.

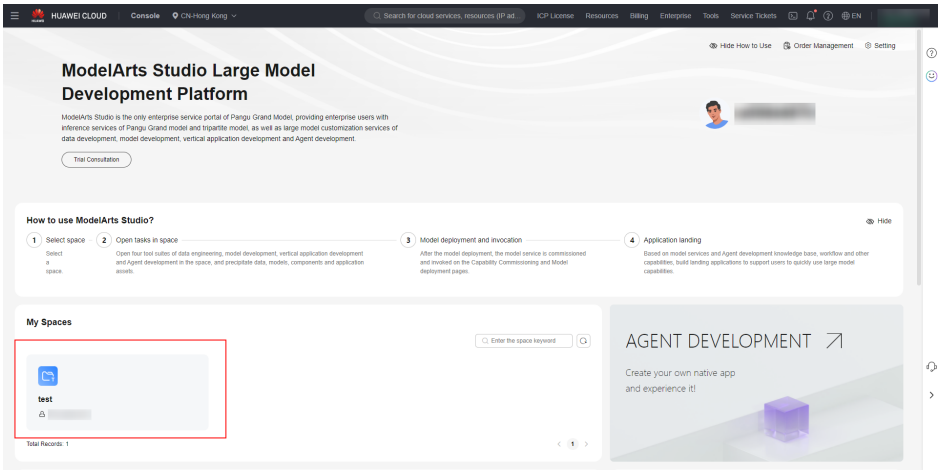
4. Click **Create Now**.
5. After a fine-tuning job is created, the **Model Training** page is displayed. You can view the job status at any time.

Full Fine-Tuning

To create a DeepSeek model full fine-tuning job, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 5-2 My Spaces



2. In the navigation pane, choose **Model Development** > **Model Training**. Click **Create Training Job** in the upper right corner.
3. On the **Create Training Job** page, set training parameters by referring to **Table 5-6**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 5-6 Parameters for full fine-tuning the DeepSeek model

Category	Training Parameter	Description
Training configuration	Select Model	You can modify the following information: • **Sources**: Select **Model Square**. • **Type**: Select **NLP** and select the base model and version (DeepSeek-V3-32K or DeepSeek-R1-32K) used for training.
	Type	Select **Supervised fine-tuning**.
	Training Objective	Select **Full fine-tuning**. • Full fine-tuning: When supervised fine-tuning is performed on a model, all parameters of the model are updated.
	Advanced Settings	During a model training job, checkpoints are used to store the model weight and status. • **Close**: After this function is disabled, checkpoints are not saved and training cannot be resumed based on checkpoints. • **Automatic**: All checkpoints during training are automatically saved. • **Custom**: A specified number of checkpoints are saved based on the settings.
Training Parameters	epochs	Dataset iterations during model training.
	learning_rate	It determines the step size at each training iteration. If the learning rate is too high, the model never converges. If the learning rate is too low, the model can take a long time to converge.
	batch_size	It controls the number of samples used in one iteration of model training. A larger batch size leads to more stable gradients but consumes more GPU memory. This may cause out-of-memory (OOM) errors due to hardware limitations and extend the training time.
	sequence_length	It specifies the maximum length of a single training data record. Data that exceeds the length will be truncated during training.
	warmup	It controls the duration of the warm-up phase relative to the entire training process. During this phase, the learning rate starts low, allowing the model to stabilize gradually. Once stabilized, the model is then trained at a preset higher learning rate, which accelerates convergence and enhances performance.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	lr_decay_ratio	It controls the learning rate to ensure the model converges more stably. The minimum learning rate will not be lower than the product of the **Learning Rate** and this parameter. If the learning rate does not decay, set this parameter to **1**.
	weight_decay	It is a regularization method that reduces the size of model parameters to prevent model overfitting and improve the model generalization capability.
	checkpoint_save_strategy	It can be set to **save_checkpoint_steps** or **save_checkpoint_epoch**. It specifies whether to save checkpoint files based on the number of iteration steps or based on the number of epochs during training.
	save_checkpoint_steps or save_checkpoint_epoch	**save_checkpoint_steps** specifies the number of training steps at which the model checkpoint file is saved during training. **save_checkpoint_epoch** specifies the number of training epochs during which the model checkpoint file is saved.
	checkpoints	During a model training job, checkpoints are used to store the model weight and status. ● **Close**: After this function is disabled, checkpoints are not saved and training cannot be resumed based on checkpoints. ● **Automatic**: All checkpoints during training are automatically saved. ● **Custom**: A specified number of checkpoints are saved based on the settings.
Data configuration	Training set	Select the training dataset.
Resource Disposition	Billing model	Billing mode of the current training job.
	Training Unit	Select the number of training units required for training the model. The minimum number of training units required for the current training is displayed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	Single-instance training units	Select the number of training units for a single instance.
	Number of instances	Select the number of instances.
Subscription reminder	Subscription reminder	After this function is enabled, the system sends SMS or email notifications to users when the task status is updated.
publish model	Enable automatic publishing	If this function is disabled, the model will be manually published to the model asset library after the training is complete. If this function is enabled, configure the visibility, model name, and description.
Basic Information	Name	Name of a training job.
	Description	Description of the training job.

NOTE

The default values of training parameters vary depending on the model. The default values displayed on the frontend page prevail.

4. Click **Create Now**.
5. After a fine-tuning job is created, the **Model Training** page is displayed. You can view the job status at any time.

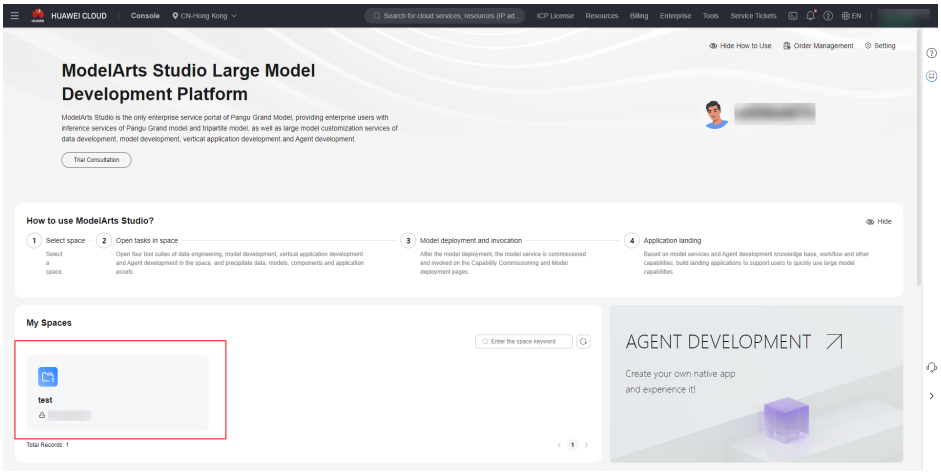
LoRA Fine-Tuning

To create a DeepSeek model LoRA fine-tuning job, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-3 My Spaces



2. In the navigation pane, choose **Model Development** > **Model Training**. Click **Create Training Job** in the upper right corner.
3. On the **Create Training Job** page, set training parameters by referring to **Table 5-7**.

Table 5-7 Parameters for DeepSeek model LoRA fine-tuning

Category	Training Parameter	Description
Training configuration	Select Model	You can modify the following information: • **Sources**: Select **Model Square**. • **Type**: Select **NLP** and select the base model and version (DeepSeek-V3-32K or DeepSeek-R1-32K) used for training.
	Type	Select **Supervised fine-tuning**.
	Training Objective	Select **LoRA fine-tuning**. • LoRA fine-tuning: During supervised fine-tuning, the pre-trained model's weight parameters are frozen. Low-rank decomposition is applied to the weight matrix in the self-attention module while retaining the original weights. During training, only the low-rank parameters are updated.
Training Parameter	epochs	Dataset iterations during model training.
	learning_rate	It determines the step size at each training iteration. If the learning rate is too high, the model never converges. If the learning rate is too low, the model can take a long time to converge.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	batch_size	It controls the number of samples used in one iteration of model training. A larger batch size leads to more stable gradients but consumes more GPU memory. This may cause out-of-memory (OOM) errors due to hardware limitations and extend the training time.
	sequence_length	It specifies the maximum length of a single training data record. Data that exceeds the length will be truncated during training.
	warmup	It controls the duration of the warm-up phase relative to the entire training process. During this phase, the learning rate starts low, allowing the model to stabilize gradually. Once stabilized, the model is then trained at a preset higher learning rate, which accelerates convergence and enhances performance.
	lr_decay_ratio	It controls the learning rate to ensure the model converges more stably. The minimum learning rate will not be lower than the product of the **Learning Rate** and this parameter. If the learning rate does not decay, set this parameter to **1**.
	weight_decay	Adds a penalty term related to the model weight size to the loss function, encouraging the model to keep small weights to prevent overfitting or overly complex models.
	Rank of the LoRA matrix	In the LoRA matrix, the rank value is used to measure the complexity and information amount of the matrix. A larger value enhances the representation capability of the model, but increases the training duration. A smaller value reduces the number of parameters and the overfitting risk.
Data configuration	Training set	Select the training dataset.
Resource Disposition	Billing model	Billing mode of the current training job.
	Training Unit	Select the number of training units required for training the model. The minimum number of training units required for the current training is displayed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	Single-instance training units	Select the number of training units for a single instance.
	Number of instances	Select the number of instances.
Subscription reminder	Subscription reminder	After this function is enabled, the system sends SMS or email notifications to users when the task status is updated.
publish model	Enable automatic publishing	If this function is disabled, the model will be manually published to the model asset library after the training is complete. If this function is enabled, configure the visibility, model name, and description.
Basic Information	Name	Name of a training job.
	Description	Description of the training job.

NOTE

The default values of training parameters vary depending on the model. The default values displayed on the frontend page prevail.

4. Click **Create Now**.
5. After a fine-tuning job is created, the **Model Training** page is displayed. You can view the job status at any time.

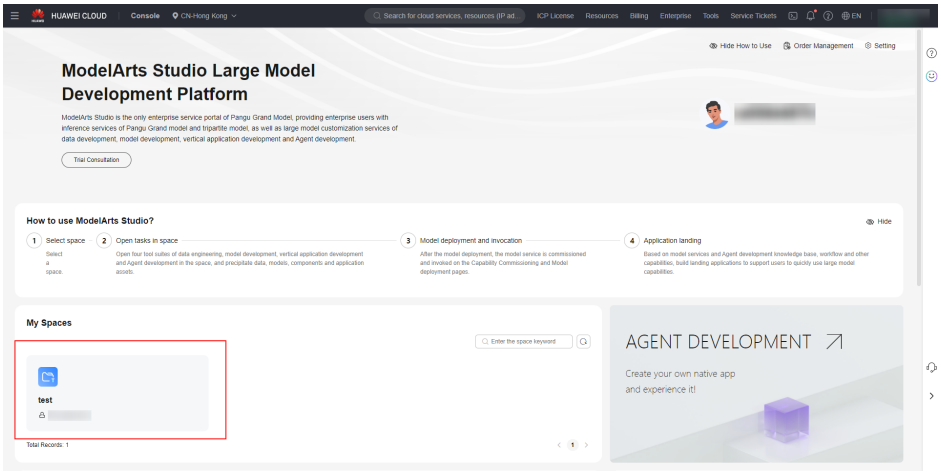
QLoRA Fine-Tuning

To create a DeepSeek model full fine-tuning job, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-4 My Spaces



2. In the navigation pane, choose **Model Development** > **Model Training**. Click **Create Training Job** in the upper right corner.
3. On the **Create Training Job** page, set training parameters by referring to **Table 5-8**.

Table 5-8 Parameters for DeepSeek model QLoRA fine-tuning

Category	Training Parameter	Description
Training configuration	Select Model	You can modify the following information: • **Sources**: Select **Model Square**. • **Type**: Select **NLP** and select the base model and version (DeepSeek-V3-32K or DeepSeek-R1-32K) used for training.
	Type	Select **Supervised fine-tuning**.
	Training Objective	Select **QLoRA fine-tuning**. • QLoRA fine-tuning: QLoRA is an efficient fine-tuning approach for LLMs. By combining quantization and low-rank adaptation, it significantly reduces GPU memory usage while maintaining model performance, making it suitable for resource-constrained environments.
Training Parameter	epochs	Dataset iterations during model training.
	learning_rate	It determines the step size at each training iteration. If the learning rate is too high, the model never converges. If the learning rate is too low, the model can take a long time to converge.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	batch_size	It controls the number of samples used in one iteration of model training. A larger batch size leads to more stable gradients but consumes more GPU memory. This may cause out-of-memory (OOM) errors due to hardware limitations and extend the training time.
	sequence_length	It specifies the maximum length of a single training data record. Data that exceeds the length will be truncated during training.
	warmup	It controls the duration of the warm-up phase relative to the entire training process. During this phase, the learning rate starts low, allowing the model to stabilize gradually. Once stabilized, the model is then trained at a preset higher learning rate, which accelerates convergence and enhances performance.
	lr_decay_ratio	It controls the learning rate to ensure the model converges more stably. The minimum learning rate will not be lower than the product of the **Learning Rate** and this parameter. If the learning rate does not decay, set this parameter to **1**.
	weight_decay	It is a regularization method that reduces the size of model parameters to prevent model overfitting and improve the model generalization capability.
	Rank of the LoRA matrix	In the LoRA matrix, the rank value is used to measure the complexity and information amount of the matrix. A larger value enhances the representation capability of the model, but increases the training duration. A smaller value reduces the number of parameters and the overfitting risk.
Data configuration	Training set	Select the training dataset.
Resource Disposition	Billing model	Billing mode of the current training job.
	Training Unit	Select the number of training units required for training the model. The minimum number of training units required for the current training is displayed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Training Parameter	Description
	Single-instance training units	Select the number of training units for a single instance.
	Number of instances	Select the number of instances.
Subscription reminder	Subscription reminder	After this function is enabled, the system sends SMS or email notifications to users when the task status is updated.
publish model	Enable automatic publishing	If this function is disabled, the model will be manually published to the model asset library after the training is complete. If this function is enabled, configure the visibility, model name, and description.
Basic Information	Name	Name of a training job.
	Description	Description of the training job.

NOTE

The default values of training parameters vary depending on the model. The default values displayed on the frontend page prevail.

4. Click **Create Now**.
5. After a fine-tuning job is created, the **Model Training** page is displayed. You can view the job status at any time.

5.3.2 Viewing the Training Status and Metrics of the DeepSeek Model

Viewing the Model Training Status

After the **model training is started, you can view the model training job status on the Model Training page.** You can click a job name to go to the details page and view the training metrics, training job details, and training logs.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 5-9 Training statuses

Status	Meaning
Initializing	The model training job is being initialized and is ready for training.
Queuing	The model training job is queuing. Please wait.
Running	The model is being trained and the training process is not complete.
Stopping	Model training is being stopped.
Stopped	Model training has been manually stopped.
Failed	An error occurred during model training. You can view logs to locate the cause of the training failure.
Completed	Model training is complete.

Viewing Training Metrics

For a training job in **Completed** state, click the job name to view the training metrics on the **Train Result** tab page.

Table 5-10 Training metric descriptions

Model	Training Metric	Description
Third-party large model	Training Loss	The training loss is a metric that measures the difference between the predicted and actual outputs of a model. In normal cases, a smaller value indicates better model performance. Generally, a normal loss curve should be monotonically decreasing. That is, as the training proceeds, the loss value decreases continuously until it converges to a smaller value.

Obtaining training logs

Click the name of a training job to view the logs generated during the training on the **Log** tab page.

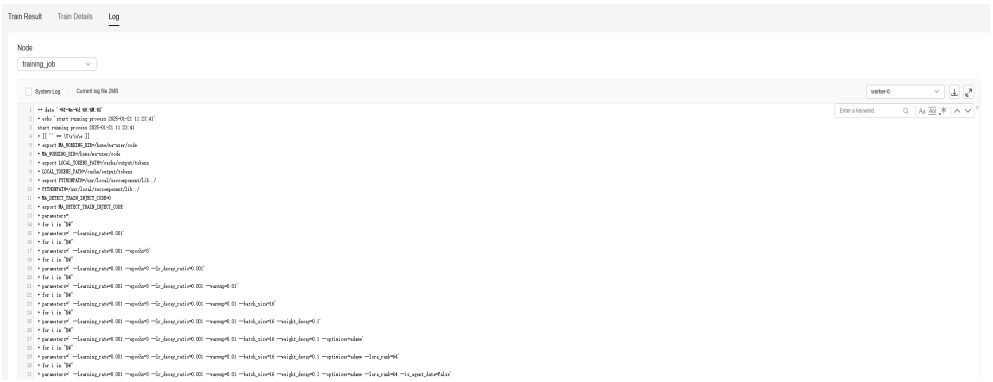
If a training job is abnormal or fails, you can locate the cause of the failure based on the training logs.

Training logs can be filtered and viewed by training phase. During distributed training, jobs are allocated to multiple worker nodes for parallel processing. Each worker node is responsible for processing a part of data or executing a specific

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

computing task. You can also filter logs by worker node (for example, worker-0 indicates the first worker node).

Figure 5-5 Obtaining training logs



5.3.3 Publishing the Trained DeepSeek Model

After a DeepSeek model is trained, perform the following steps to publish it:

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.
2. In the navigation pane, choose **Model Development** > **Model Training**. Click the model name to go to the job details page.
3. Click the **Train Result** tab and click **Release**.
4. Enter the asset name and description, select the visibility, and click **OK** to publish the model.

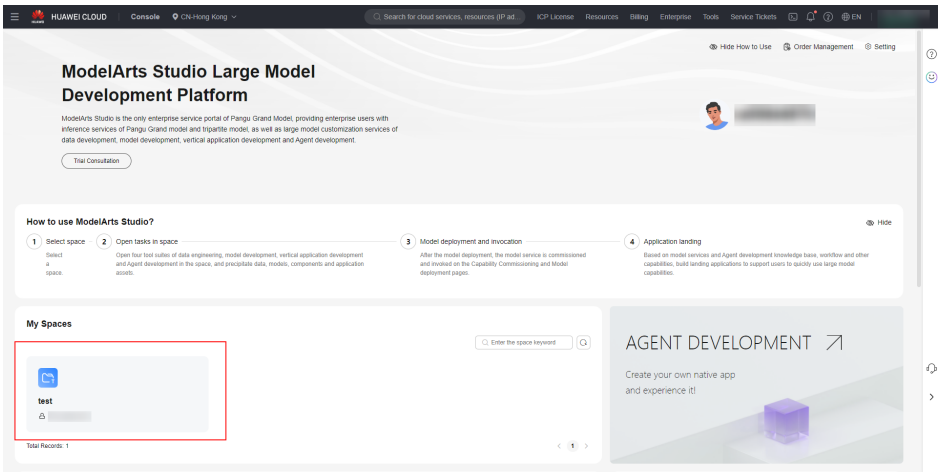
   The published model is displayed as a model asset in the **Models** list of the **Workspace Assets** page.

5.3.4 Managing DeepSeek Model Training Jobs

In the training job list, a job creator can clone (copy a training job), retry (retrain a job), and delete a job.

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

Figure 5-6 My Spaces



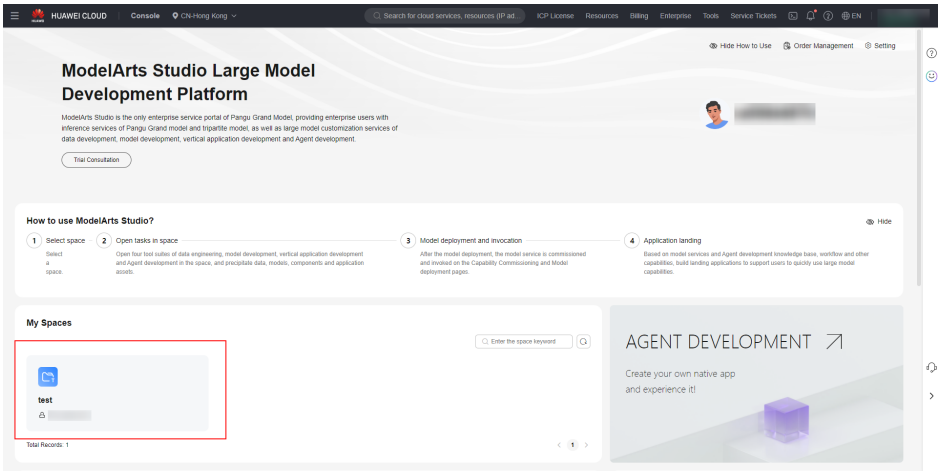
2. In the navigation pane, choose **Model Development** > **Model Training** and perform the following operations:
   - Clone: Click **Clone** in the **Operation** column to clone the current training job.
   - Retry: Click **Retry** in the **Operation** column to edit the node that fails to be executed and retry the training of the node.
   - Delete: Click **Delete** in the **Operation** column to delete unnecessary training jobs.

5.4 Compressing a DeepSeek Model

Before model deployment, perform model compression, so less video memory is needed during inference and inference performance is improved. By compressing a third-party large model, the storage size of the model can be reduced, power consumption can be decreased, and computational speed can be improved.

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

Figure 5-7 My Spaces



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane, choose **Model Development** > **Model Compression**. Click **Create** in the upper right corner. Create a model compression task by referring to **Table 5-11**.

Table 5-11 Parameters for creating a model compression task

Category	Parameter Name	Description
Compression Configuration	Model Source	Select **Third-party large model**.
	Base Model	Select the model to be compressed. You can use a model from assets or jobs.
	Compression Policy	Select **INT8**.
Resource Disposition	Resources Pool	Select the resource pool used by the model compression task.
Subscription reminder	Subscription reminder	After this function is enabled, the system sends SMS or email notifications to users when the task status is updated.
publish model	Enable automatic publishing	If this function is disabled, the model will be manually published to the model asset library after the training is complete. If this function is enabled, configure the visibility, model name, and description.
Basic Information	Task name	Name of a model compression task.
	Compressed model name	Name of the compressed model. (This configuration item is available only when **Enable automatic publishing** is disabled.)
	Description	Name of a model compression task.

3. After setting the parameters, click **Create Now**.

5.5 Deploying a DeepSeek Model

After the model training is complete, you can deploy the model.

1. Log in to ModelArts Studio and access the required workspace.
2. In the navigation pane, choose **Model Development** > **Model Deployment**. Click **Create Deployment Task** in the upper right corner.
3. In the **Select Model** dialog box, set **Sources** to **Model Square** and **Type** to **NLP**, select a model, and click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4. On the **Create a deployment** page, set deployment parameters by referring to **Table 5-12** and start model deployment.

Table 5-12 Parameters for deploying a third-party large model

Category	Parameter	Description
Deployment Configuration	Select Model	You can modify the following information: • **Sources**: Select **Model Square**. • **Type**: Select **NLP** and select the model and version to be deployed.
	Deployment Mode	**Deployment at the online**: Algorithms are deployed in the resource pool provided by the platform.
	Deployed_Model	Unique identifier of the inference service when the inference service is called through the inference API of the V2 version.
Safety guardrail	Open and agree to authorize	Ensures the security of model calling.
	Version Selection	Currently, only the basic edition of the security guardrail is supported, which is built-in with default content moderation rules.
Resource Disposition	Billing model	Free for a limited time
	Number of instances	Set the number of instances required by the model to be deployed. It is recommended that the number of instances to be deployed at a time be less than or equal to 10. Otherwise, traffic limiting may be triggered and the deployment may fail.
Subscription reminder	Subscription reminder	After this function is enabled, the system sends SMS or email notifications to users when the task status is updated.
Basic Information	Service Name	Set the name of the deployment task.
	Description (Optional)	Set the description of the deployment task.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

5. After setting the deployment parameters, click **Deploy Now**.

5.6 Evaluating a DeepSeek Model

5.6.1 Creating a DeepSeek Model Evaluation Dataset

A model with excellent performance must possess strong generalization capabilities. This means that the model should perform well not only on the provided data (training data) but also on unseen data. To achieve this goal, model evaluation is indispensable. When collecting evaluation datasets, ensure that the datasets are independent and random to guarantee that the collected data can represent a sample of real-world data. This helps to avoid biasing the evaluation result, thereby more accurately reflecting the performance of the model in different scenarios. By using the evaluation dataset to evaluate a model, developers can understand the advantages and disadvantages of the model and find the optimization direction.

The **DeepSeek model** supports both manual and automatic evaluation modes.

- Manual evaluation: The answers generated by the model are evaluated based on the manually created evaluation dataset and evaluation metrics. During evaluation, the model answers need to be manually scored based on the created evaluation items. After the evaluation is complete, an evaluation report is generated based on the scoring result.
- Automatic evaluation: There are two types of evaluation rules: rule-based and LLM-based.
  - The answers generated by the model are automatically evaluated based on rules (similarity/accuracy). You can use the professional datasets preset in the evaluation template or customize evaluation datasets. For details about the supported preset datasets, see **Table 5-13**.
  - The model is used to automatically score the generated result of the evaluation model, which is applicable to open or complex Q&A scenarios. There are grading mode and comparison mode.

Table 5-13 Preset evaluation sets

Type	Dataset	Description
General knowledge and skills	Common knowledge	Evaluate the mastery of basic knowledge and information about daily life, including history, geography, and culture.
	Mathematical capability	Evaluate the model's capability of solving mathematical problems, including arithmetic operations, algebraic equation solving, and geometric shape analysis.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

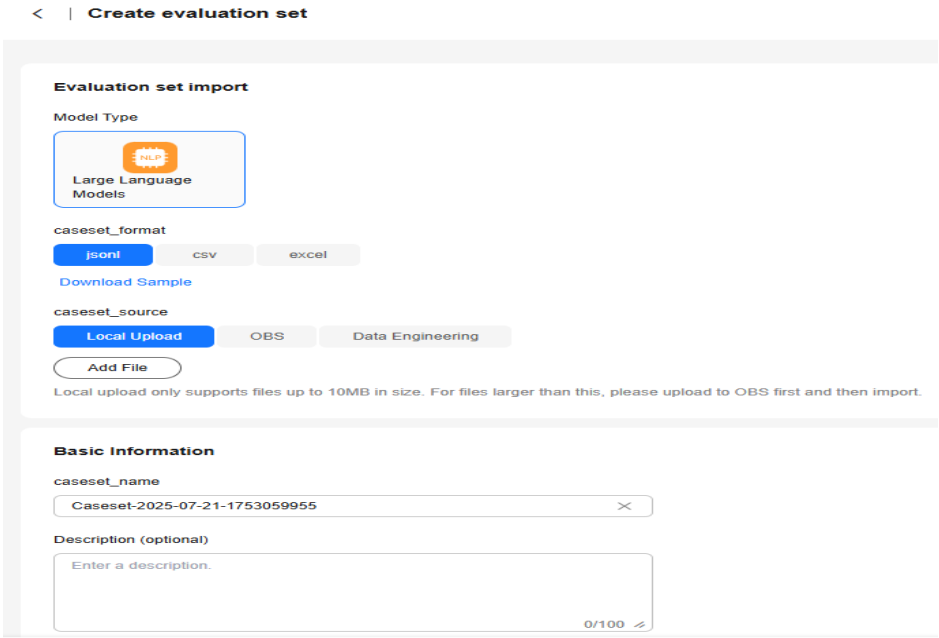
Type	Dataset	Description
	Logical reasoning	Evaluate the capability of the evaluation model to perform reasonable inference and analysis based on known information, involving multiple logical processes of thinking such as deduction and induction.
	Long text	Evaluate the ability of a large model to read and understand long text materials, including extracting key information and summarizing the content.
	Chinese capability	Evaluate the advanced knowledge and inference capabilities of large models in the Chinese language and cultural background.
Domain knowledge	Finance	Evaluate the capabilities of large models in the financial field, including quick understanding and interpretation of complex financial concepts, risk prediction and data analysis, investment suggestions, and financial decision-making support.

Importing or Uploading Evaluation Datasets from a Local PC, OBS files, and Data Engineering Datasets

1. Log in to ModelArts Studio and access the required workspace.
2. In the navigation pane, choose **Evaluation Center** > **Evaluation Set Management**. Click **Create evaluation set** in the upper right corner.
3. On the **Create evaluation set** page, select the required file content, file format, and import source, and select a path to upload the data file.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-8 Importing an evaluation dataset



Creating a Dataset Using Data Engineering

The procedure for creating an evaluation dataset is the same as that for creating a training dataset. This section describes only the procedure. For details, see **Using Data Engineering to Build a DeepSeek Model Dataset**.

1. Log in to ModelArts Studio and access the required workspace.
2. In the navigation pane, choose **Data Engineering** > **Data Acquisition**. On the **Import Task** page, click **Create Import Job** in the upper right corner.
3. On the **Create Import Job** page, select the required file content, file format, and import source, and click **Select Path** to upload the data file.

   Table 5-14 lists the formats supported by the NLP model evaluation dataset.

Table 5-14 Evaluation dataset format

Model Type	Evaluation Dataset Format
DeepSeek model	Text - Single Round QA - jsonl
DeepSeek model	Other - Customization -csv

NOTE

In the NLP multi-turn Q&A evaluation scenario, choose **Data Engineering** > **Data Acquisition**. Click **Create Import Job**. Set **Dataset Type** to **Other**. Import the multi-turn Q&A evaluation set in CSV format.

No., number of rounds, question, and reference reply

1,1,Hello. What can I do for you?

1,2,Please introduce Huawei Cloud products, including but not limited to computing, storage, and network products.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4. After the data file is uploaded, enter the dataset name and description, and click **Create Now**.
5. In the navigation pane, choose **Data Engineering** > **Data Publishing** > **Publish Task**. On the displayed page, click **Create Data Publish Task** in the upper right corner.
6. On the **Create Data Publish Task** page, select a dataset modality and a dataset file.
7. Click **Next**, select a format, enter the name, select the dataset visibility, and click **OK**.

NOTE

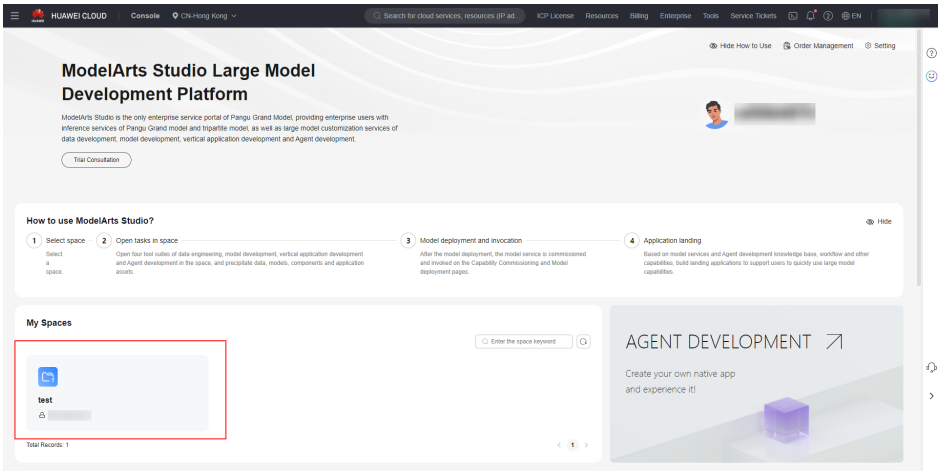
To evaluate a Pangu model, you need to publish the dataset in Pangu format.

5.6.2 (Optional) Creating an API

The API management module can manage APIs of the evaluation services. You can select an evaluation service when creating an evaluation job. You can create, edit, and delete APIs.

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

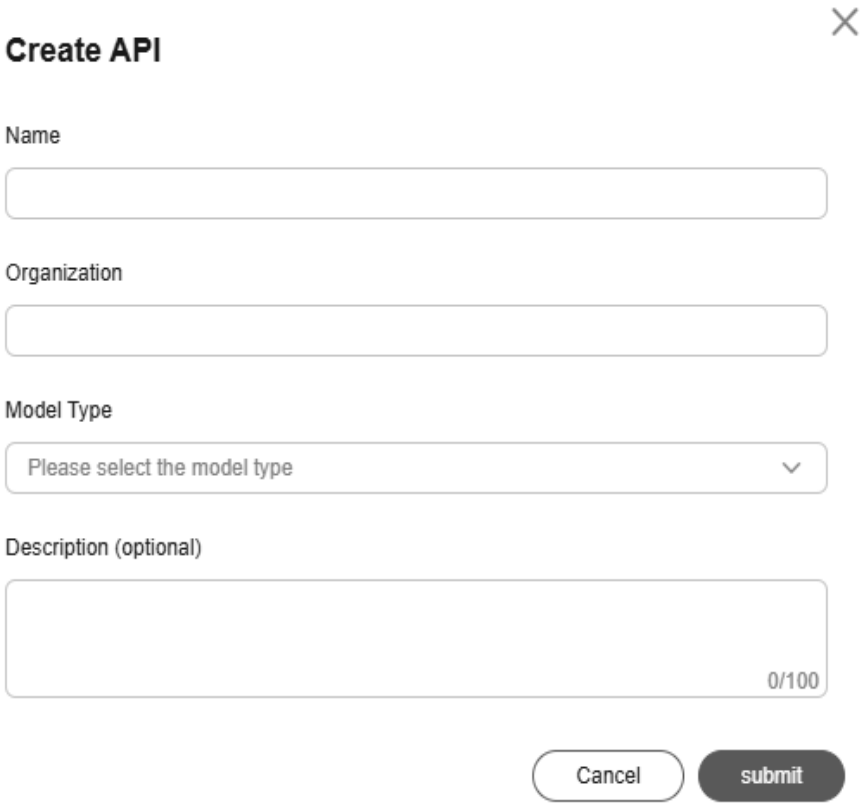
Figure 5-9 My Spaces



2. In the navigation pane, choose **Evaluation Center** > **API Management**. Click **Create API** in the upper right corner.
3. Set the API name and organization, select a model type, and click **submit**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-10 Creating an API



4. After an API is created, add versions on the right to define the API information. You can add multiple versions.
5. Click **New Version** and configure the **model version**, **Concurrency Level**, **QPM**, **URL**, **Param**, **Body**, and **Response**.
   a. **model version**: user-defined.
   b. **Concurrency Level**: Set it based on the model performance. The value ranges from 1 to 32.
   c. **QPM**: Set it based on the model performance. The value ranges from 1 to 5000.
   d. **URL**: Set a model inference address.
   e. **Param**: You can set the parameters based on the API requirements. You can define the header and request parameters.
   f. **Body**: inference request body. You can customize the body or select the preset OpenAI or TGI structure.
   g. **Response**: defines how to obtain the content in the response body in JsonPath format. The JsonPath syntax is used to extract required data from the JSON field in the response body.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

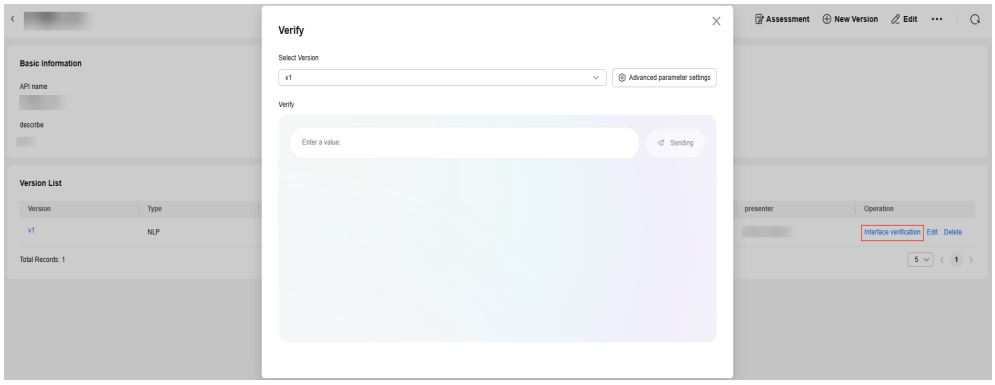
Figure 5-11 Adding a version for an API



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

6. After creating an API version, you need to verify whether the API is available. Only verified APIs can be evaluated. On the **API Management** page, click the name of the created API task, click **Interface verification**, enter a question, and check whether the API response content is displayed.

Figure 5-12 Interface verification



Creating an API Service

In this example, we use a deployed and properly running Pangu NLP model to create an API service.

1. On the **Model Development** > **Model Deployment** page, find the deployed NLP model, click the service name, and copy the API URL on the details page.
2. In the navigation pane, choose **Evaluation Center** > **API Management**. Click **Create API** in the upper right corner. Define the API name and organization structure, and set the model type to NLP.
3. After an API is created, add versions on the right to define the API information. You can add multiple versions. Click **New Version** and set parameters based on **Table 5-15**.

Table 5-15 Configuring the API service

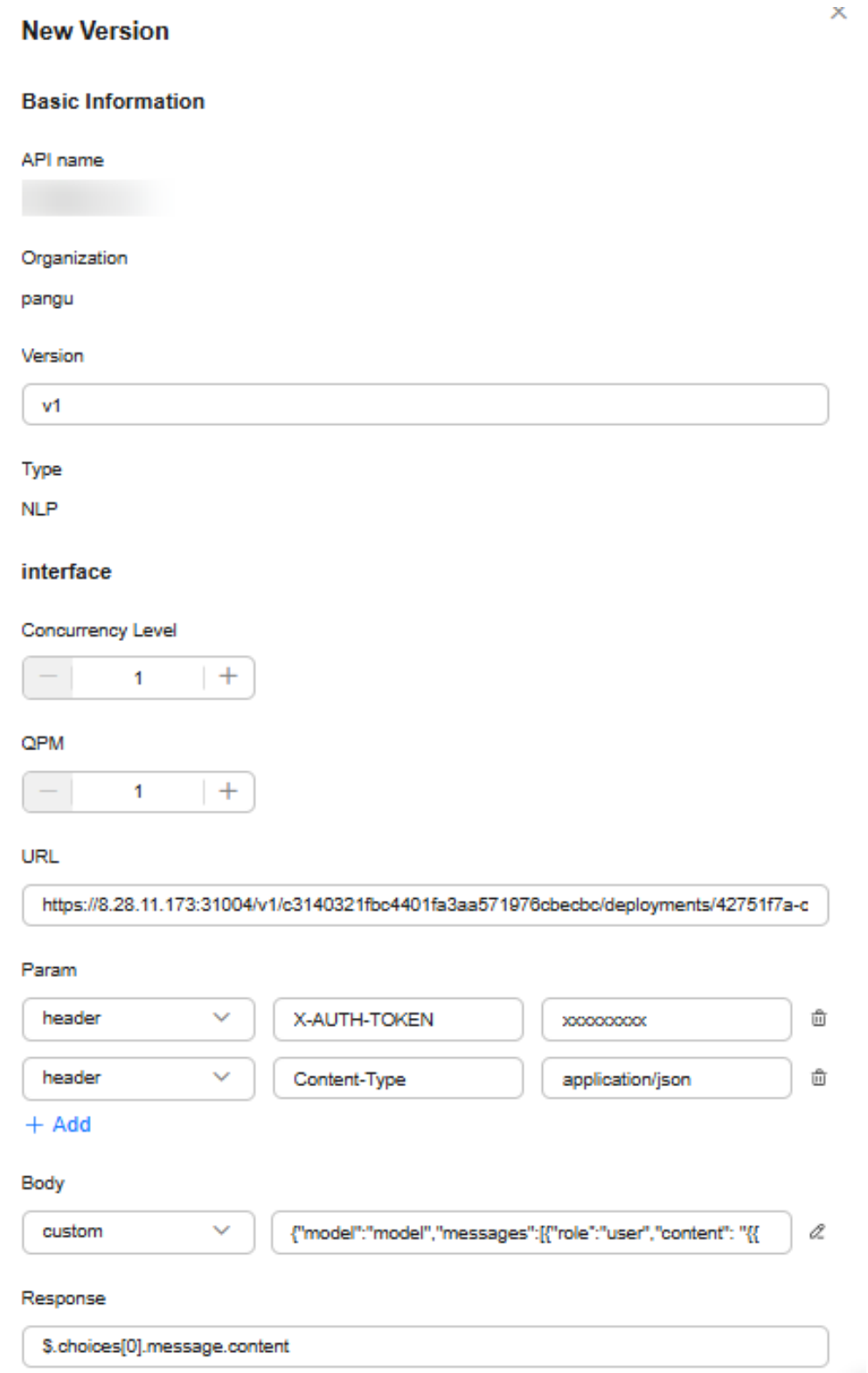
Parameter	Description
model version	The value can be customized.
Concurrency Level	Set this parameter to **1**.
QPM	Set this parameter to **1**.
URL	Enter the obtained API URL.
Param	Set two parameters of the **header** type. Parameter 1: The parameter name is **X-AUTH-TOKEN**, and the parameter value is the token value. For details about how to obtain the token, see "Calling REST APIs" > "Authentication" in *API Reference*. Parameter 2: The parameter name is **Content-Type** and the parameter value is **application/json**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter	Description
Body	Select **custom** and enter the following content: {"model": "model","messages": [{"role": "user","content": "{{message}}"}],"max_tokens": 1024, "temperature": 0.7,"top_p": 0.95,"presence_penalty": 0.5,"frequency_penalty": 0.5,"stream": false}
Response	Enter the following content: $.choices[0].message.content

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-13 Configuring the API service



5.6.3 Creating a DeepSeek Model Evaluation Job

Before creating an NLP model evaluation job, ensure that the operations in **Creating a DeepSeek Model Evaluation Dataset** have been completed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

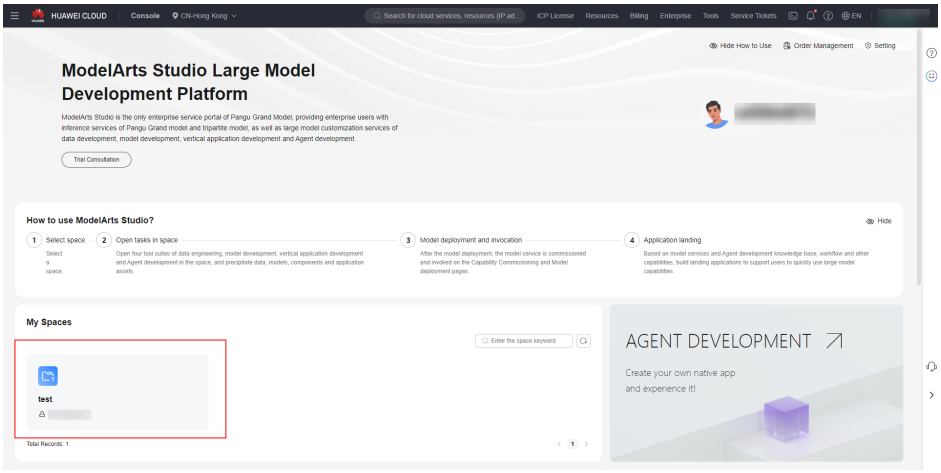
Pre-trained NLP models cannot be evaluated.

Creating a Rule-based Automatic Evaluation Task for a DeepSeek Model

To create an automatic evaluation job for an NLP model, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 5-14 My Spaces

2. In the navigation pane, choose **Evaluation Center** > **Evaluation task**. Click **Create automatic task** in the upper right corner.
3. On the **Create automatic Evaluation task** page, set deployment parameters by referring to **Table 5-16**.

   Table 5-16 Parameters for a rule-based automatic evaluation task of a DeepSeek model

Category	Parameter Name	Description
selection service	Model Type	Select **Large Language Models**.

Category	Parameter Name	Description
	Service source	Two options are available: **Deploying services** and **External services**. A maximum of 10 models can be evaluated at a time. ● **Deploying services**: Select a model deployed on ModelArts Studio for evaluation. ● **External services**: Access external models through APIs for evaluation. When selecting **External services**, you need to enter the API name, API address, request body, and response body of the external model. – The request body can be in OpenAI, TGI, or custom format. The openai format is a large model request format developed and standardized by OpenAI. The tgi format is a large model request format launched by the HuggingFace team. – The response body of the API must be entered based on the JsonPath syntax requirements. The JsonPath syntax is used to extract required data from the JSON field in the response body.
Evaluation Configurations	Evaluation Rules	Select **Rule-based**: Automatic scoring is performed based on rules. That is, scoring is performed based on similarity or accuracy, and the difference between the model's prediction and the labeled data is compared. It is applicable to standard multiple-choice questions or simple Q&A scenarios.
	Evaluation Dataset	● Preset evaluation dataset: Use a preset professional dataset for evaluation. ● **Single review set**: You can specify evaluation metrics (F1 score, accuracy, BLEU, and rouge) and upload an evaluation dataset for evaluation. If you select **Single review set**, you need to upload the dataset used for the evaluation.
	Storage location of measurement results	Path for storing the model evaluation result.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
Basic Information	Task Name	Enter the evaluation job name.
	Description	Enter the evaluation job description.

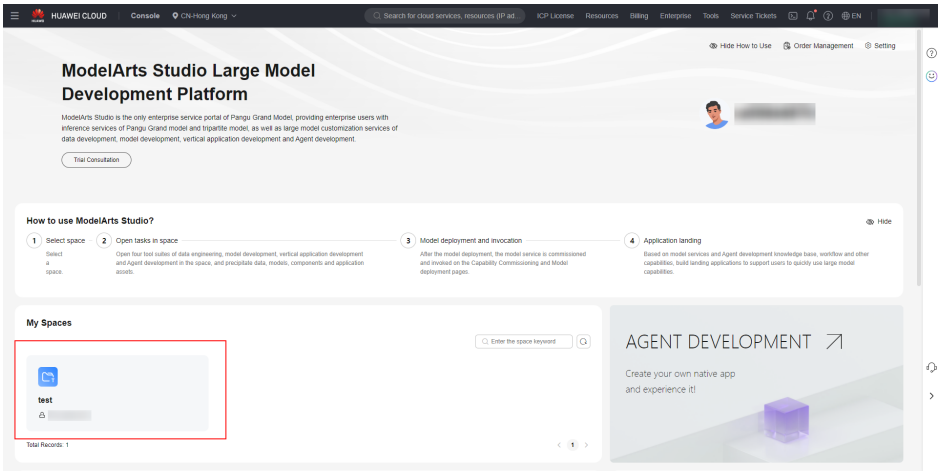
4. After setting the parameters, click **Create Now**. The **Evaluation Task** > **Automatic evaluation** page is displayed.
5. When the status is **Completed**, you can click **Evaluation Report** in the **Operation** column to view the model evaluation result, including the detailed score and evaluation details.

Creating an LLM-based Automatic Evaluation Task for a DeepSeek Model

To create an automatic evaluation job for a DeepSeek model, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 5-15 My Spaces



2. In the navigation pane, choose **Evaluation Center** > **Evaluation task**. Click **Create automatic task** in the upper right corner.
3. On the **Create automatic Evaluation task** page, set parameters by referring to **Table 5-17**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 5-17 Parameters for an LLM-based automatic evaluation task of a DeepSeek model

Category	Parameter Name	Description
selection service	Model Type	Select **Large Language Models**.
	Service source	Two options are available: **Deploying services** and **External services**. A maximum of 10 models can be evaluated at a time. • **Deploying services**: Select a model deployed on ModelArts Studio for evaluation. • **External services**: Access external models through APIs for evaluation. When selecting **External services**, you need to enter the API name, API address, request body, and response body of the external model. – The request body can be in OpenAI, TGI, or custom format. The openai format is a large model request format developed and standardized by OpenAI. The tgi format is a large model request format launched by the HuggingFace team. – The response body of the API must be entered based on the JsonPath syntax requirements. The JsonPath syntax is used to extract required data from the JSON field in the response body.
Evaluation Configurations	Evaluation Rules	**Based on large models**: Use large models with stronger capabilities to automatically score the generated results of the evaluated models. This approach is applicable to open or complex Q&A scenarios.
	Select mode	• **grading model**: The referee model automatically scores the model inference result based on the configured scoring criteria. • **Comparison mode**: The referee model compares the performance of two models on each question. The comparison result can be win, lose, or tie. In comparison mode, two services must be selected as the service source. By default, the first service is selected as the benchmark model.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
	Scoring Prompt Template	• In grading mode, the default value is score_prompt. The prompt contains the standard reply of the current scenario. The prompt is input to the referee model for scoring in the scoring phase. • In comparison mode, the default value is arena_prompt. The prompt contains the standard reply of the current scenario, which is used by the referee model to compare the advantages and disadvantages of two services. During this process, you can modify metrics such as the metric evaluation dimension in the **Variables** area on the right. You can modify the scoring metrics and steps.
	Evaluation Dataset	Select the dataset to be evaluated. In the NLP multi-turn Q&A scenario, only automatic evaluation based on large models is supported. You can select the multi-turn Q&A evaluation dataset.
	Storage location of measurement results	Path for storing the model evaluation result.
Referee configuration	Referee Model	You can select a deployed service or an external service.
	Scoring rules	Scoring rules can be customized. The referee model scores or compares model results based on the configured rules.
Basic Information	Task Name	Enter the evaluation job name.
	Description	Enter the evaluation job description.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

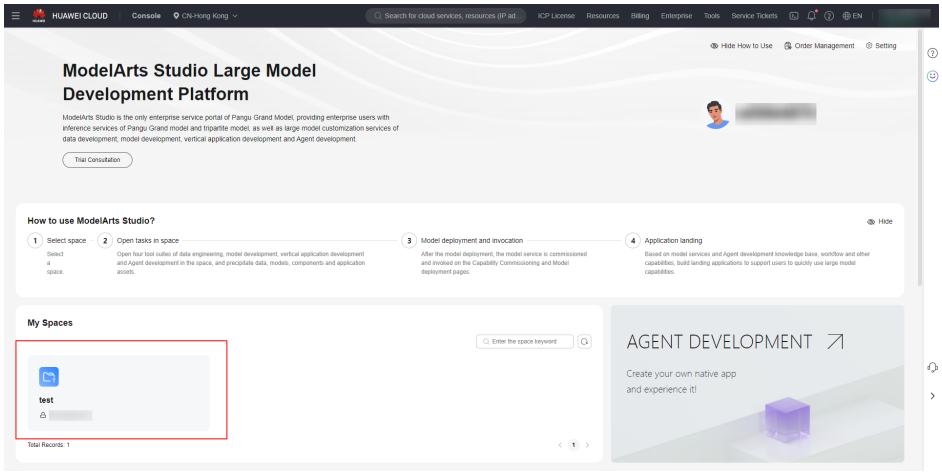
4. After setting the parameters, click **Create Now**. The **Evaluation Task** > **Automatic evaluation** page is displayed.
5. When the status is Completed, you can click **Evaluation Report** in the **Operation** column to view the model evaluation result, including the detailed score and evaluation details.

Creating a Manual Evaluation Job for a DeepSeek Model

To create a manual evaluation job for an NLP model, perform the following steps:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 5-16 My Spaces



2. In the navigation pane, choose **Evaluation Center** > **Evaluation task**. Click **Create manual task** in the upper right corner.
3. On the **Create manual Evaluation task** page, set parameters by referring to **Table 5-18**.

Table 5-18 Parameters of creating a manual evaluation job for a DeepSeek model

Category	Parameter Name	Description
selection service	Model Type	Select **Large Language Models**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
	Service source	Two options are available: **Deploying services** and **External services**. A maximum of 10 models can be evaluated at a time. • **Deploying services**: Select a model deployed on ModelArts Studio for evaluation. • **External services**: Access external models through APIs for evaluation. When selecting **External services**, you need to enter the API name, API address, request body, and response body of the external model. – The request body can be in OpenAI, TGI, or custom format. The openai format is a large model request format developed and standardized by OpenAI. The tgi format is a large model request format launched by the HuggingFace team. – The response body of the API must be entered based on the JsonPath syntax requirements. The JsonPath syntax is used to extract required data from the JSON field in the response body.
Evaluation Configurations	Evaluation Indicators	You can customize evaluation metrics and fill in evaluation standards.
	Evaluation Dataset	Evaluation dataset.
	Storage location of measurement results	Path for storing the model evaluation result.
Basic Information	Task Name	Enter the evaluation job name.
	Description	Enter the evaluation job description.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

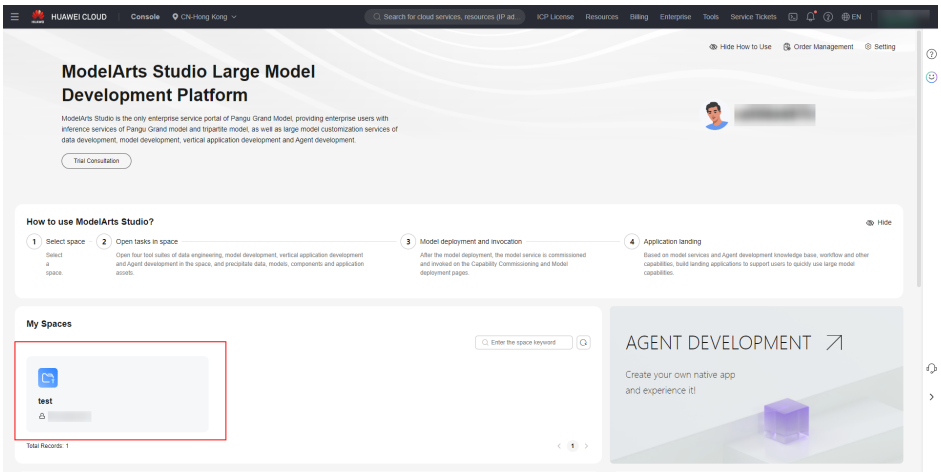
4. After setting the parameters, click **Create Now**. The **Evaluation Task** > **Manual evaluation** page is displayed.
5. When the status is **To be evaluated**, you can click **Online Evaluation** in the **Operation** column to go to the evaluation page.
6. Score the evaluation effect area as prompted. After all data is evaluated, click **Submit**.
7. On the **Manual evaluation** tab page, check that the status of the evaluation job is **Completed**. Click **Assessment report** in the **Operation** column to view the model evaluation result.

5.6.4 Viewing the DeepSeek Model Evaluation Report

After an evaluation job is created, you can view the task report. The procedure is as follows:

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 5-17 My Spaces



2. In the navigation pane, choose **Model Development** > **Model Evaluation** > **Job Management**.
3. Click **Assessment report** in the **Operation** column. On the displayed page, you can view the basic information and overview of the evaluation job.

   For details about each evaluation metric, see **DeepSeek Model Evaluation Metrics**.
4. Export the evaluation report.
   a. On the **Evaluation Report** > **Evaluation Details** page, click **Export**, select the report to be exported, and click **OK**.
   b. Click **Export Record** on the right to view the exported task ID. Click **Download** in the **Operation** column to download the evaluation report to the local PC.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

DeepSeek Model Evaluation Metrics

The DeepSeek Model supports automatic evaluation and manual evaluation. For details about the metrics, see **Table 5-19**, **Table 5-20**, and **Table 5-21**.

Table 5-19 Automatic evaluation metrics of the DeepSeek model (not using preset evaluation datasets)

Evaluation Metric (Automatic Evaluation - Evaluation Template Not Used)	Description
F1_SCORE	Harmonic mean of the precision and recall. A larger value indicates better model performance.
BLEU-1	Matching degree between the sentence generated by the model and the actual sentence at the single-word level. A larger value indicates better model performance.
BLEU-2	Matching degree between the sentence generated by the model and the actual sentence at the phrase level. A larger value indicates better model performance.
BLEU-4	Weighted average accuracy of the model generation result and actual sentences. A larger value indicates better model performance.
ROUGE-1	Similarity between the sentence generated by the model and the actual sentence at the single-word level. A larger value indicates better model performance.
ROUGE-2	Similarity between the sentence generated by the model and the actual sentence at the two-word level. A larger value indicates better model performance.
ROUGE-L	Similarity between the sentence generated by the model and the actual sentence at the longest common subsequence (LCS) level. A larger value indicates better model performance.
PRECISION	Accuracy of Q&A matching, that is, the accuracy of sentences generated by the model compared with actual sentences. A larger value indicates better model performance.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 5-20 Automatic evaluation metrics of the NLP model (using preset evaluation datasets)

Evaluation Metric (Automatic Evaluation-Evaluation Template Used)	Description
Score	The score of each dataset is the pass rate of the model in the current dataset. If there are multiple datasets in the evaluation capability items, the weighted average pass rate is calculated based on the data volume.
Comprehensive capability	The comprehensive capability is the weighted average of the pass rates of all datasets.

Table 5-21 Manual evaluation metrics of the DeepSeek model

Evaluation Metric (Manual Evaluation)	Description
Accuracy	The answer generated by the model is correct and there is no factual error.
average	The model calculates the average score of the generated sentence and the actual sentence based on the evaluation metric.
goodcase	The model calculates the proportion of test cases whose score is 5 after the generated sentence and the actual sentence are compared based on the evaluation metric.
badcase	The model calculates the proportion of test cases whose score is less than 1 after the generated sentence and the actual sentence are compared based on the evaluation metric.
Custom metric	Custom metrics, such as usability, logic, and security.

5.6.5 Managing DeepSeek Model Evaluation Jobs

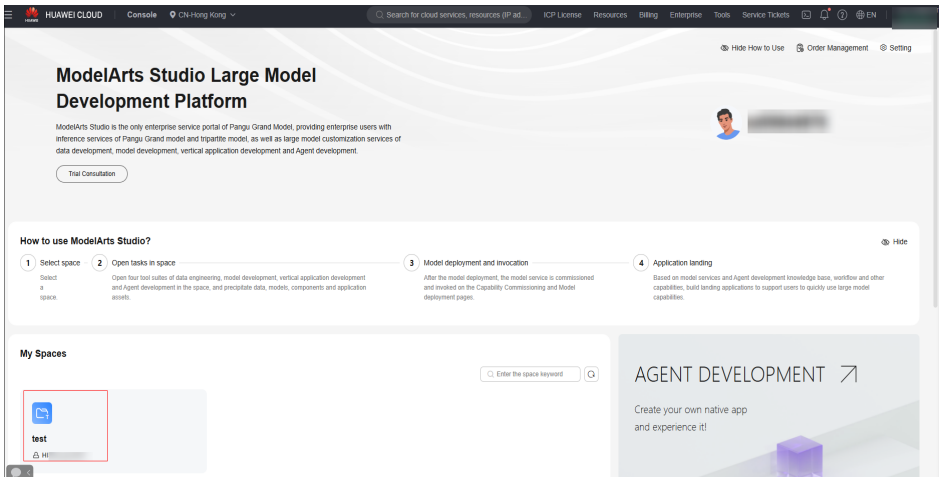
Managing Evaluation Jobs

In the evaluation job list, the job creator can clone (copy an evaluation job), start (restart a job), and delete a job.

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-18 My Spaces



2. In the navigation pane, choose **Evaluation Center** > **Evaluation Task**. You can perform the following operations:
   - Clone: Click **Clone** in the **Operation** column to copy an evaluation job in the Completed state.
   - Start: Click **Start** in the **Operation** column to restart the evaluation job that fails to be executed.
   - Delete: Click **Delete** in the **Operation** column to delete an unnecessary evaluation job in the Completed state.

     NOTE

     Deleting an evaluation job is a high-risk operation. Before deleting a job, ensure that it is no longer needed.

5.7 Calling a DeepSeek Model

Obtaining the API Request URL of the Third-Party Model

The API request URL consists of the API domain name and API URI.

After the model is deployed, you can call the following API to call the third-party model.

Table 5-22 Text completion API (example)

API Type	API URI
Deepseek-xxx	{api-url}/generate

To obtain the API domain name, choose **Model Development** > **Model Deployment** on ModelArts Studio. The detailed procedure is as follows:

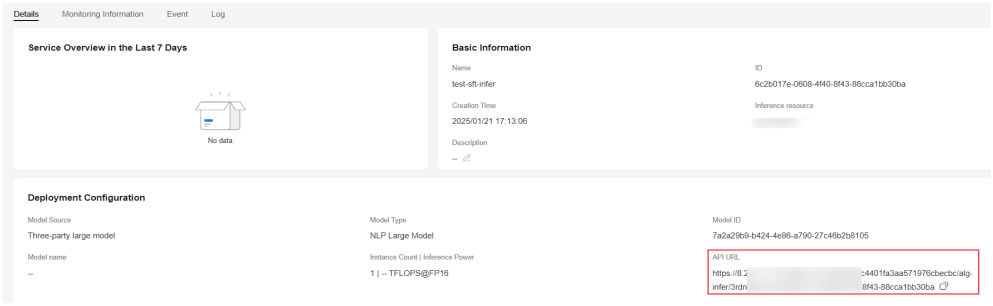
1. Log in to ModelArts Studio and access the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane on the left, choose **Model Development** > **Model Deployment**. Click a model name and obtain the API domain name on the details page.

   Only models in the **Running** state can be called.

   Figure 5-19 Obtaining the API domain name of a DeepSeek model



3. After obtaining the API domain name, combine it with the API URI to form a complete API request URL.

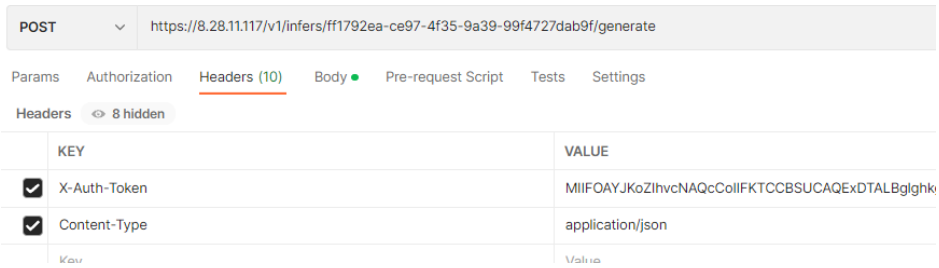
Obtaining a Token

Obtain the token by following the instructions provided in section "Calling REST APIs for Authentication" in *API Reference*.

Using APIs to Call a DeepSeek Model

1. Create a POST request in Postman and enter the API request URL.
2. Set two request header parameters by referring to **Figure 5-20**.
   - **KEY**: **Content-Type**; **VALUE**: **application/json**
   - **KEY**: **X-Auth-Token**; **VALUE**: the token value obtained in this section

   Figure 5-20 Entering a third-party model API

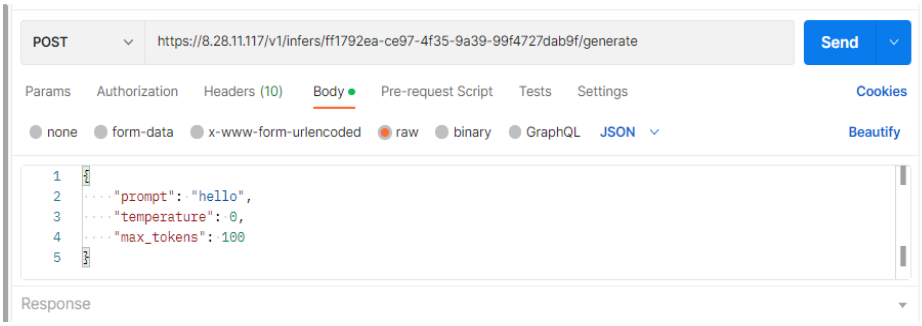


3. Click **Body**, select **raw**, refer to the following code, and enter the request body.

```
{
   "prompt": " Hello,"
   "temperature": 0,
   "max_tokens": 100
}
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 5-21 Entering the request body of the third-party model



4. Click **Send** on Postman to send the request.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

6 Developing a Prompt Engineering Project

6.1 What Is Prompt Engineering?

Introduction

Prompt engineering is the relatively new discipline of developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and in a wide variety of research fields. Prompt engineering skills help us better understand the capabilities and limitations of large language models (LLMs).

Prompt engineering is not just about designing and developing prompts. It encompasses a wide range of skills and techniques that are useful for interacting and developing with LLMs. It is an important skill for understanding the capabilities of and working effectively with LLMs. You can use prompt engineering to improve safety of LLMs and develop new capabilities, enabling you to augment LLMs with domain knowledge and external tools.

Elements of a Prompt

You can achieve a lot with simple prompts, but the quality of the results depends on how much information you provide it and how well-crafted the prompt is. A prompt can contain information like the instructions or question you are passing to the model and can include other details such as additional context, inputs, or examples. You can use these elements to instruct the model more effectively to improve the quality of results. A prompt contains any of the following elements:

- **Instruction**: a specific task or instruction you want the model to perform, for example, to summarize, extract, or generate something
- **Context**: external information or additional context that can steer the model to better responses
- **Input data**: the input or question that you are interested in finding a response for

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- **Output indicator**: the type or format of the output

You do not need all four elements for a prompt, and the format depends on the task at hand.

Process for Using Prompt Engineering

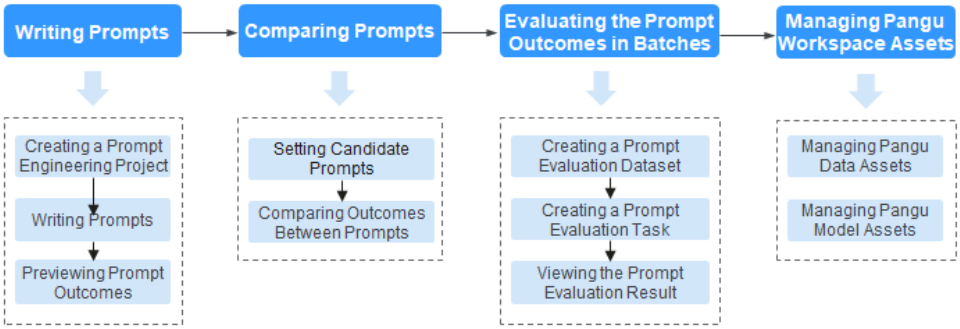
The large model development platform ModelArts Studio helps you write, compare, evaluate, save, and manage prompts.

Table 6-1 Function descriptions

Function	Description
Prompt engineering project management	The prompt engineering platform manages prompt engineering projects. Each project is tailored to a particular scenario or optimization requirement. You can optimize, compare, and evaluate prompts in a prompt engineering project. You can create, query, modify, and delete prompt engineering projects.
Prompt writing	You can edit prompts, define prompt variables, view the outcomes generated from prompts, and track the optimization history.
Candidate prompt management	You can manage candidate prompts that have been optimized. You can save a maximum of nine candidate prompts per prompt engineering project. These prompts can be compared and evaluated later.
Prompt comparison	You can select two candidate prompts to compare their content and parameters. You can set the same variable values for them to compare their outcomes.
Prompt evaluation	You can manage prompt evaluation tasks, for example, creating, querying, modifying, and deleting tasks. You can create an evaluation task, select candidate prompts and variable sets, select the evaluation method, and evaluate the results generated from those prompts.
Prompt management	You can save satisfactory prompts as candidates and query or delete prompts.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-1 Process for using prompt engineering



6.2 Obtaining a Prompt Template

The platform provides prompt templates for multiple task scenarios to help you better use the capabilities of large models and steer models towards generating more accurate and targeted outputs, thereby improving the performance of models in specific tasks.

Before creating a prompt engineering project, you can use the preset prompt template or reconstruct the prompt template. You can also create one.

To obtain prompt templates, choose **Agent Dev** > **Prompt Engineering** > **Prompt template**.

6.3 Writing Prompts

6.3.1 Creating a Prompt Engineering Project

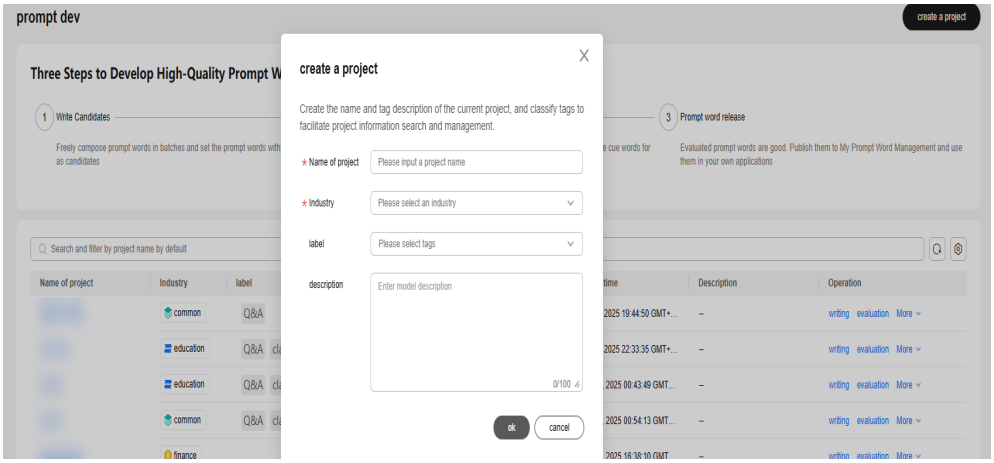
Well-designed and tailored prompts can greatly improve the quality of the LLM's responses. With prompt engineering, you can discover the most effective ways to utilize prompts to enhance the LLM's performance across different tasks.

Before writing prompts, you need to create a prompt engineering project to facilitate unified prompt management.

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Agent Dev**. On the displayed page, choose **Prompt Engineering** > **Prompt dev**, and click **create a project**.
3. Enter the project name and description, and select the industry and label. Click **ok**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-2 Creating a prompt engineering project



6.3.2 Writing Prompts

A prompt is a piece of text that is used to guide an AI model to generate a specific output. A prompt should contain key information about the target task or domain, such as the theme, style, and format.

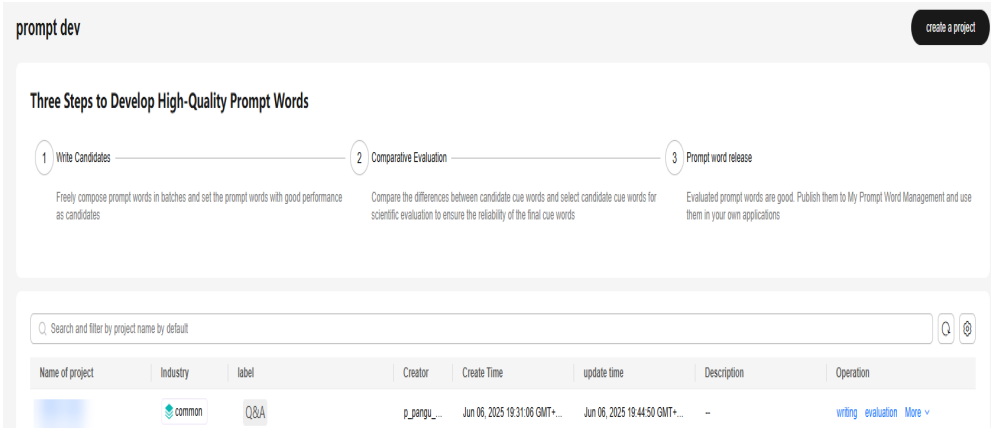
When writing a prompt, you can set variables in it. Specifically, use placeholders {{}} in a prompt as an indication of dynamic information. In this way, the model can generate varying results based on different situations, enhancing its flexibility and adaptability. For example, the prompt can be set to "You are a travel assistant. You need to introduce the local customs of a travel place to travelers. Please introduce the local customs of {{location}}." When evaluating the prompt outcomes, you can replace the values of {{location}} in batches to obtain the model answer, improving the evaluation efficiency.

When writing a prompt, you can adjust the setting of LLM parameters such as **Temperature**, **Top P**, and **Max Tokens** to control how the LLM generates text. They can affect the quality and diversity of the model's generated results.

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Agent Dev**. On the displayed page, choose **Prompt Engineering** > **Prompt dev**.
3. In the prompt engineering project list, locate the desired project and click **writing** in the **Operation** column.

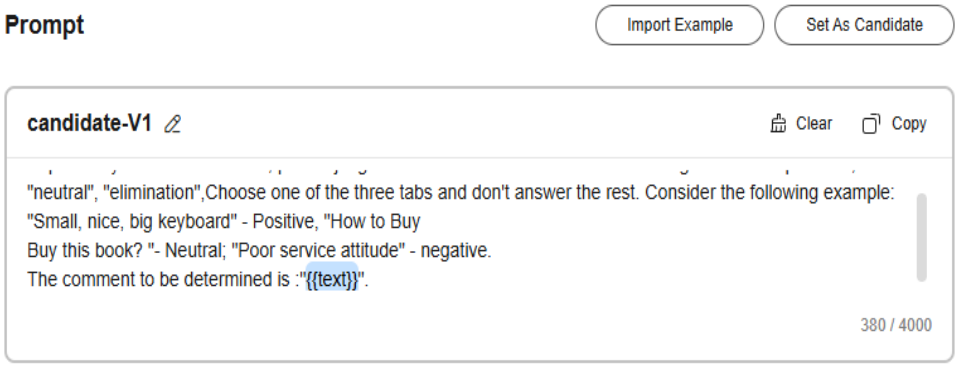
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-3 Prompt Engineering



4. Enter the prompt text in the prompt writing area. You can insert multiple variables into the prompt. The variables must be identified by placeholders {{}}.

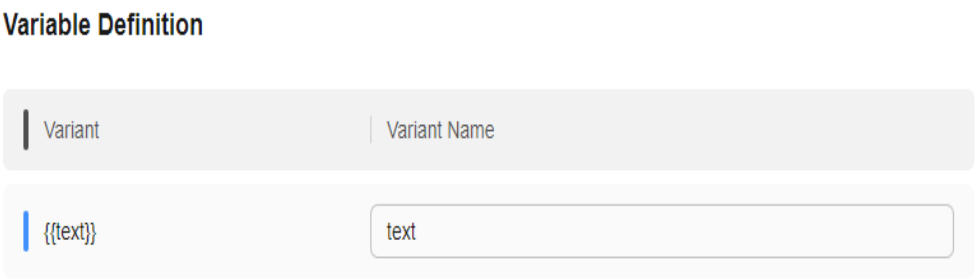
Figure 6-4 Writing prompts



5. Click **Confirm**. The platform automatically identifies the inserted variables. The variables recognized in the prompt are displayed in the **Variable Definition** area.

   You can modify the variable name, for example, add remarks for the model to better understand the function of the variable.

Figure 6-5 Variable Definition



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

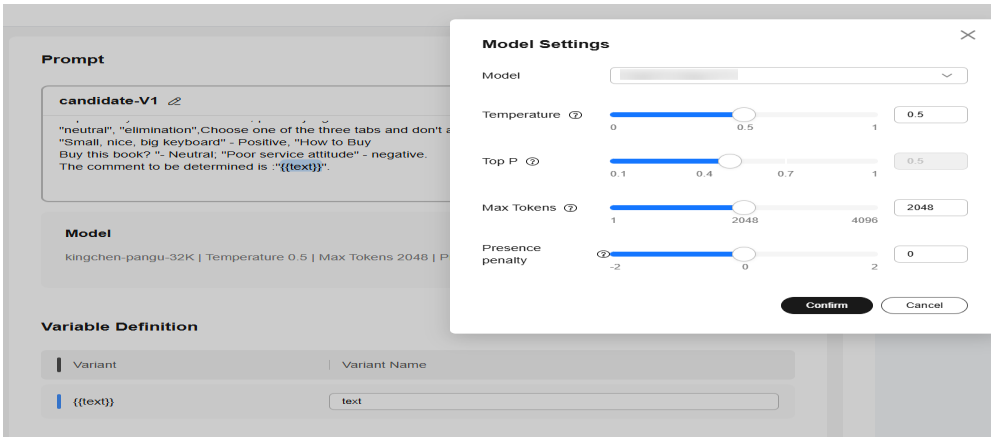
NOTE

The **Variable Definition** area displays the variables defined in the entire project, including variables associated with the candidate prompts. For details about the operations related to candidate prompts, see **Setting Candidate Prompts**.

A maximum of 20 variables can be defined in a prompt engineering project.

6. In the **Model** area, click **Settings** to set the desired model and its parameters.

Figure 6-6 Model Settings



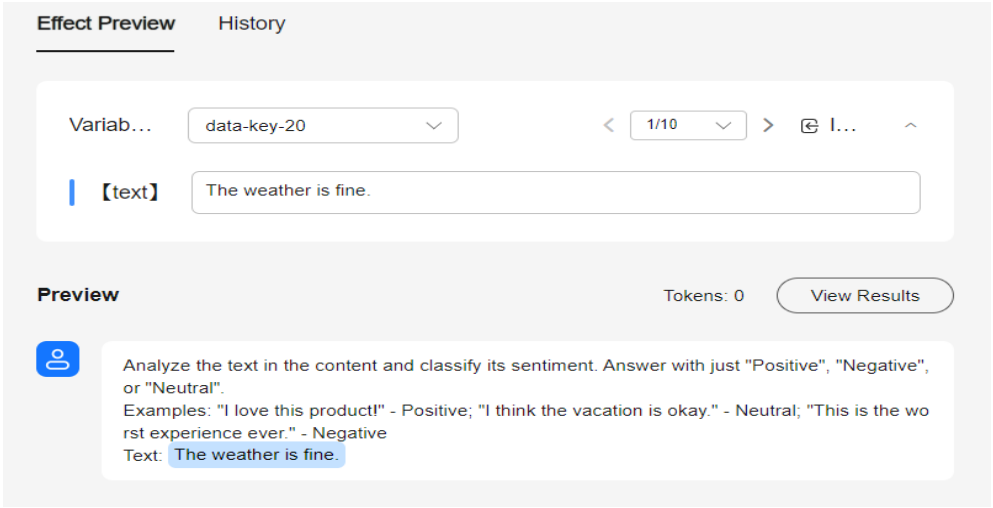
6.3.3 Previewing Prompt Outcomes

Set values for prompt variables, feed the model with complete prompts, and observe the outcomes from different prompts.

1. In the **Variable Input** area of the **writing** page, enter the variables in the text box.

   After variables are defined, the preview area will automatically show the complete prompt. You can also select an existing variable set and enter variable values. A variable set is an Excel file. Each row of data contains variable values to be entered. You can use the Import function to upload a variable set.

Figure 6-7 Effect Preview



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. Click **View Results** to view the model's answer. You can adjust the prompt text and variables based on the previewed outcome.

6.4 Comparing Outcomes Between Prompts

6.4.1 Setting Candidate Prompts

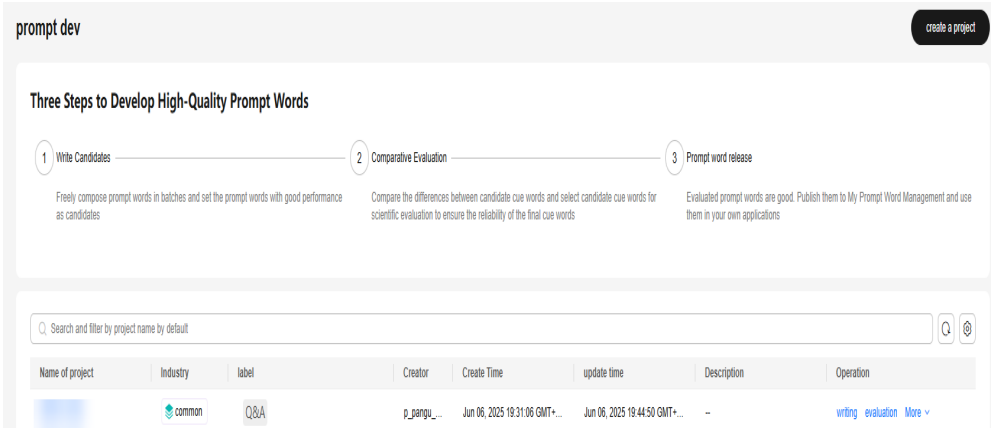
Set a prompt that yields favorable results as a candidate prompt and compare the outcomes of different candidate prompts.

NOTE

A maximum of nine candidate prompts can be added for each prompt engineering project. If the number of candidate prompts reaches this maximum, delete previously selected prompts before you can add new ones.

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Agent Dev**. On the displayed page, choose **Prompt Engineering** > **Prompt dev**.
3. In the prompt engineering project list, locate the desired project and click **writing** in the **Operation** column.

Figure 6-8 Prompt Engineering

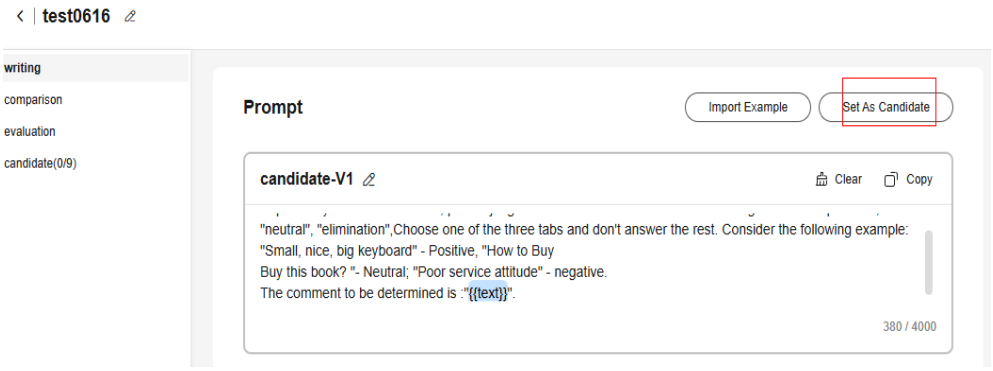


4. In the **Prompt** area, click **Set As Candidate** in the upper left corner to set the current prompt as a candidate.

   The candidate prompts are saved under the **candidate** menu in the navigation pane.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-9 Set As Candidate

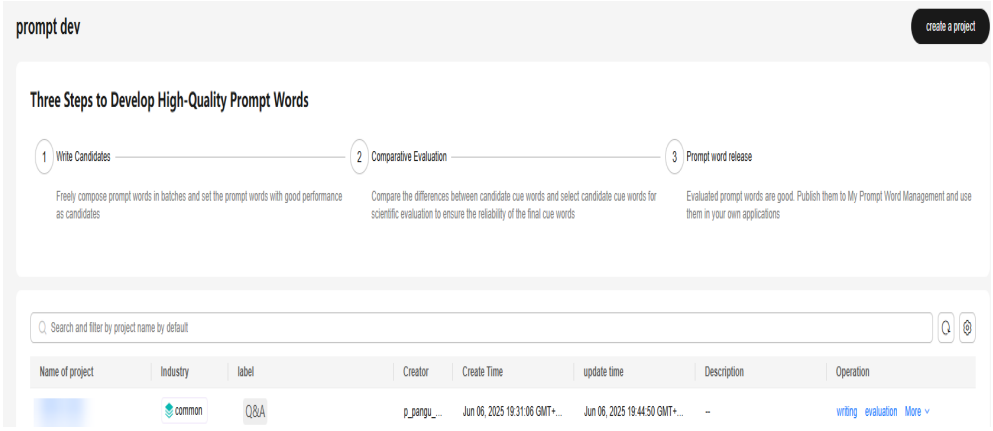


6.4.2 Comparing Outcomes Between Prompts

Perform horizontal comparison of prompt outcomes between different candidate prompts to see their differences.

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Agent Dev**. On the displayed page, choose **Prompt Engineering** > **Prompt dev**.
3. In the prompt engineering project list, locate the desired project and click **writing** in the **Operation** column.

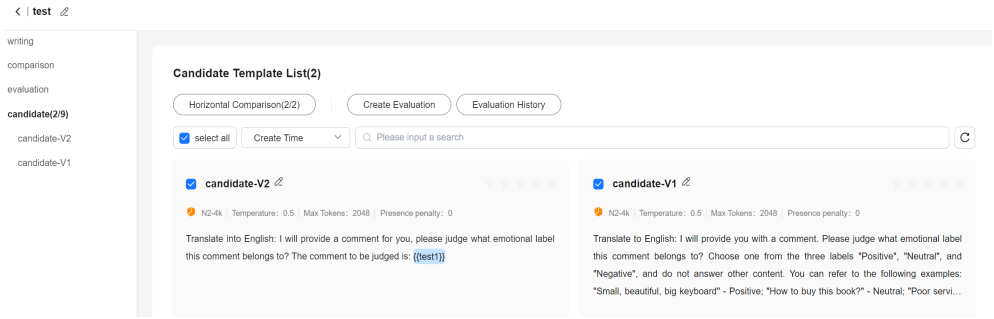
Figure 6-10 Prompt Engineering



4. On the **writing** page, choose **candidate** from the navigation pane. In the candidate list, select the prompts for horizontal comparison and click **Horizontal Comparison**.

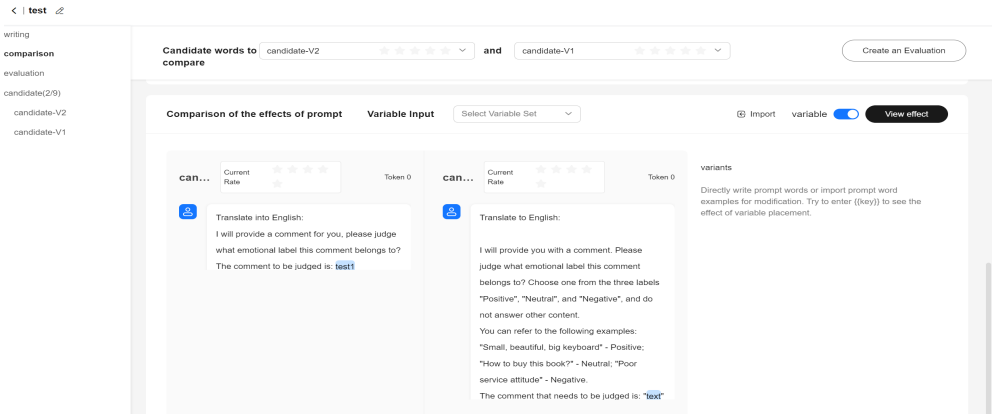
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-11 Horizontal Comparison



5. In the horizontal comparison page, scroll down to the **Comparison of the effects of prompt** area. Enter the same variable values, and compare the outcomes of the two prompts.

Figure 6-12 Prompt outcome comparison



6.5 Evaluating the Prompt Outcomes in Batches

6.5.1 Creating a Prompt Evaluation Dataset

Before evaluating the prompt outcomes, upload a prompt variable data file to create an evaluation dataset.

Prompt variables are placeholders within a prompt that can be replaced with specific value. They are used for generating different content based on different scenarios or user inputs. The variable name can be any text, pertaining to the meaning or function of the variable.

Constraints on the Prompt Evaluation Dataset

- The file to be uploaded must be in XLSX format.
- The number of rows in the dataset must range from 10 to 50.
- The table headers in the dataset must be unique, and the number of table headers must be less than 20.
- The length of a single text data record cannot exceed 1,000 characters.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

> **NOTE**
>
> The platform verifies the dataset file you upload according to the preceding constraints during creation.

The following shows a reference format and example.

Table 6-2 Reference format

key1	key2	key3	result
Group 1 - v1	Group 1 - v2	Group 1 - v3	Group 1 - r
Group 2 - v1	Group 2 - v2	Group 2 - v3	Group 2 - r
Group 3 - v1	Group 3 - v2	Group 3 - v3	Group 3 - r

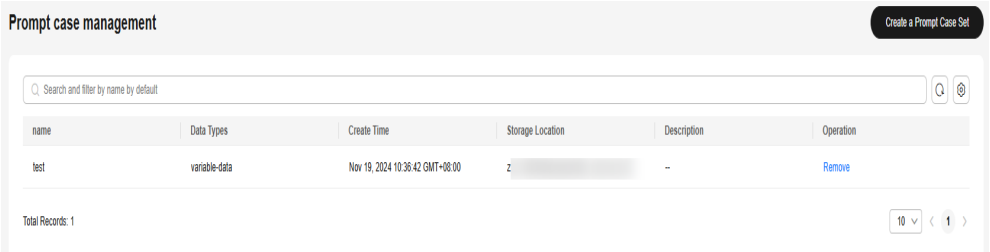
Table 6-3 Example

comment	result
How to correct misspelled words?	Neutral
This effect is not good.	Negative
I like this style.	Positive
Are you worried about something?	Neutral

Creating a Prompt Evaluation Dataset

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Agent Dev**. On the displayed page, choose **Prompt Engineering** > **Dataset Management**. On the displayed page, click **Create a Prompt Case Set** in the upper right corner.

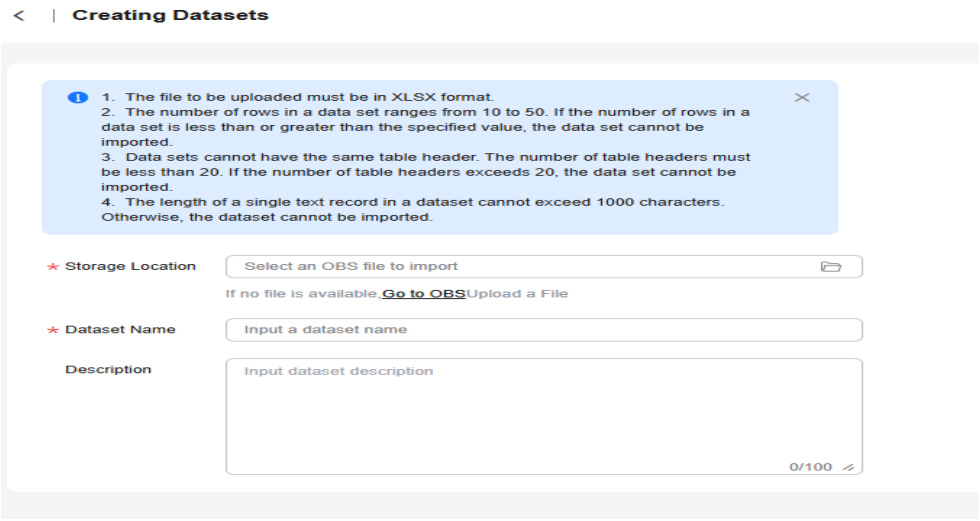
   Figure 6-13 Prompt case management

3. Upload the dataset on the **Creating Datasets** page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-14 Creating a prompt evaluation dataset

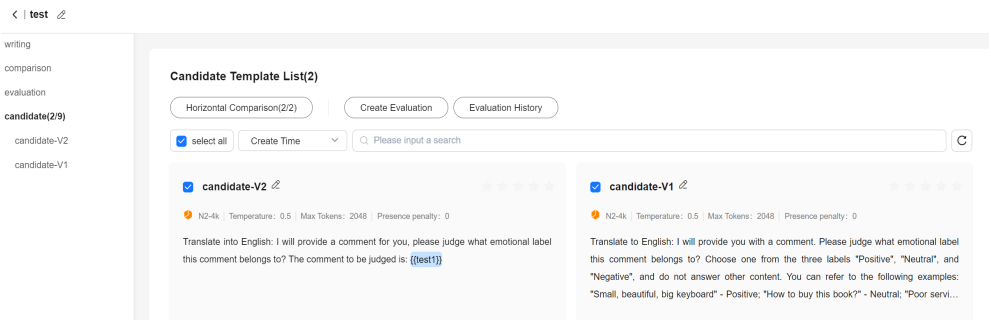


6.5.2 Creating a Prompt Evaluation Task

Create an evaluation task to batch evaluate candidate prompts. The procedure is as follows:

1. Log in to ModelArts Studio and access a workspace.
2. In the navigation pane, choose **Agent Dev**. On the displayed page, choose **Prompt Engineering** > **Prompt dev**.
3. In the prompt engineering project list, locate the desired project and click **writing** in the **Operation** column.
4. On the **writing** page, choose **candidate** from the navigation pane. In the candidate list, select the prompts for horizontal comparison and click **Create Evaluation**.

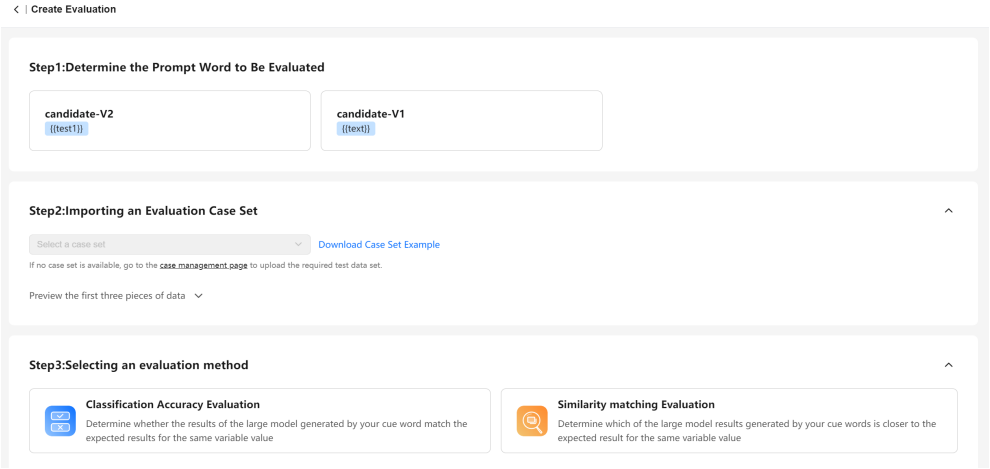
Figure 6-15 Create Evaluation



5. Select the variable set and evaluation method.
   - **Evaluation case set**: The platform assembles the prompt to be evaluated and the variables in the selected dataset into a complete prompt. The model can generate a corresponding outcome in response to this prompt.
   - **Evaluation method**: Using the selected evaluation method, the platform compares the model's generated result with the expected one, and then offers the corresponding score.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 6-16 Creating a prompt evaluation task

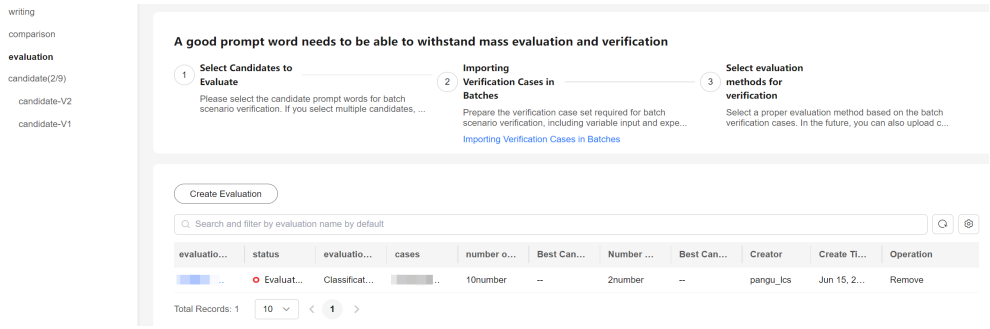


6. Click **Confirm**. The evaluation task is automatically executed.

6.5.3 Viewing the Prompt Evaluation Result

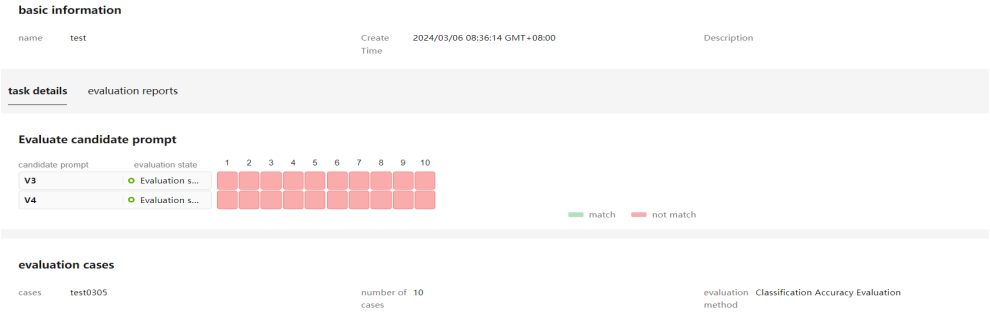
1. On the details page of the created evaluation task, view the status and report of the evaluation.

Figure 6-17 Viewing the status of a prompt evaluation task



2. Click an evaluation name to go to the evaluation task details page. You can view the evaluation progress. For example, there are 10 evaluation cases in **Figure 6-18**, among which 8 have been evaluated and the remaining two are to be evaluated.

Figure 6-18 Viewing the evaluation progress



3. After the evaluation is complete, view the evaluation result of each data record.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

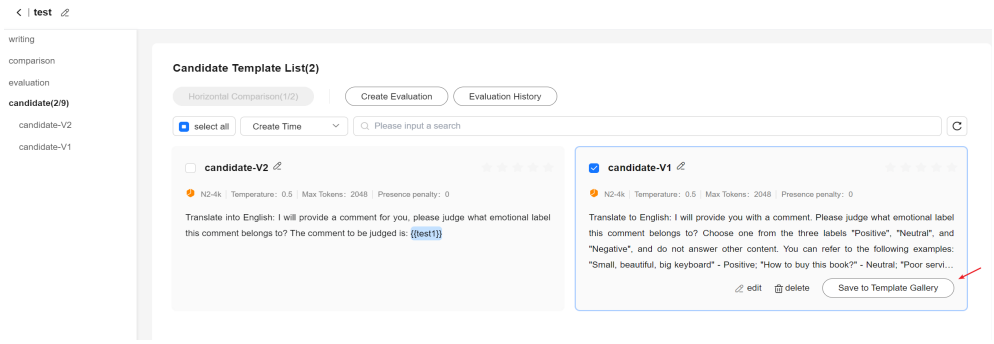
In the result, "expected result" indicates the expected answer preset for the variable value (question), and "generated result" indicates the answer provided by the model. You can determine which the prompt is better by comparing the expected result with the generated result.

6.6 Publishing Prompts

After conducting a **horizontal comparison of different prompts** and **batch evaluating their outcomes**, if you identify high-quality prompts, you can publish them to the prompt template gallery.

1. On the **Candidate Template List** page, select a good prompt and click **Save to Template Gallery**.

   Figure 6-19 Save to Template Gallery



2. On the **Agent App Dev** > **Prompt Engineering** > **Prompt template** page, view the published prompts.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

7 Developing an Agent

7.1 Agent Development Platform Overview

7.1.1 Platform Overview

An agent is an AI entity that can independently execute tasks based on user instructions or preset rules. An agent generally has certain intelligence and decision-making capabilities, and can simulate human thinking and behavior patterns to automatically complete complex operations such as information retrieval, data analysis, and task scheduling. Agents are widely used in scenarios such as customer service, automated office, and data analysis. Agents continuously improve their performance through learning and optimization to provide efficient and convenient services for users.

The agent development platform provides a complete set of tools and frameworks for creating, configuring, and managing agents. The platform supports zero-code/low-code hybrid development. You can orchestrate workflows (integrating nodes such as the LLM, knowledge retrieval, and plug-ins) on the visualized GUI to quickly configure capabilities such as intent recognition, task routing, and code execution, implementing efficient development and large-scale deployment of agents.

Zero code: Easy creation of agents

- Intelligent prompt generation: The platform provides an intelligent prompt generation function. Even if you do not have a foundation in large models, you can quickly generate suitable prompts to easily guide agents to complete tasks.
- Rich skill configuration: The platform provides various skill configuration options, covering knowledge bases, plug-ins, and workflows. Adding skills is as simple as building blocks. You only need to click and select skills to enable the agent to master new skills, such as accessing knowledge bases in specific domains to become experts in these domains.
- Quick creation and running: The entire creation process is very simple. You can bring an agent online in a few steps. From conception to actual application, the time is greatly shortened, allowing zero-code users to quickly have their own intelligent assistants.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Low code: flexible setup of complex workflows

- Visualized workflow design: The GUI encapsulates various functions into nodes, such as the large model node and intent recognition node. Developers can drag and connect nodes like a puzzle to build complex workflows. The logic is clear and visible.
- Code enhancement flexibility: Code can be written between nodes, providing more customization space for developers with programming experience. For example, you can add several lines of data processing code to the code node to implement special data conversion and meet personalized requirements.

Why Agent Development Platform?

- **Comprehensive plug-in integration capabilities**

  The agent development platform continuously expands the capability boundaries of agents through a rich plug-in ecosystem. The platform presets dozens of practical plug-ins, such as file processing and code interpreter, for out-of-the-box use. It also provides open capabilities for creating custom plug-ins. Developers can encapsulate internal APIs, third-party tools, or self-developed functions into plug-ins and enable agents to invoke the plug-ins with simple parameter configuration.

- **Flexible knowledge base management**

  The platform provides an intelligent knowledge base system, which allows developers to build dedicated data asset libraries. Local documents (PDF, Doc, and PPT formats), web page data, and structured databases can be quickly imported to the knowledge base. Th agent automatically learns the data and creates indexes to implement accurate knowledge retrieval and content generation. For example, after an industry report is uploaded, the agent can automatically generate a competitive product analysis report.

- **Modular workflow design**

  The platform uses the zero-code and low-code visualized development mode, allowing developers to quickly build complex task flows by dragging and dropping nodes. It provides nearly 100 standard components, such as natural language processing, logic judgment, and code execution, to meet the development requirements from simple Q&A to multi-step tasks (for example, building an intelligent customer service workflow that includes requirement analysis, solution design, and cost accounting).

- **Full-link agent development support**

  The platform provides a complete development and debugging system, covering code compilation, plug-in testing, and workflow verification. Developers can observe each step of the agent execution in real time, view the plug-in calling result and inference process, and quickly locate problems through hierarchical log analysis (for example, tracing the dialog logic and data calling path when the intelligent customer service handles user complaints).

- **Flexible MCP service**

  MCP uses the universal standard language to supply tools and data through the MCP server (one-off development and unlimited connections). In this way, AI agents can communicate with thousands of external tools and data more efficiently and conveniently.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Platform Architecture

The agent development platform provides one-stop AI application building capabilities. The platform provides multiple capability combinations to support Agent development scenarios with different complexities.

Basic general capabilities

The agent development platform can be used by multiple tenants and managed in a unified manner. Session management and historical dialog records can be viewed. Agents can be released as APIs.

Core functional modules

- Workstation: The workstation allows you to create zero-code apps, orchestrate apps through workflows, create custom plug-ins through plugin management and connect them to apps, create MCP services, and create and manage knowledge bases.
- App treasure box: The app treasure box contains the AppGallery, workflow market, and external apps, and supports various out-of-the-box apps.
- Plug-in market: More than 10 types of plug-ins are preset in the plug-in market, such as the file processing plug-in and code interpreter, which are out-of-the-box.
- MCP service market: Multiple external MCP APIs are preset for users to easily invoke.

Applicable scenarios

The agent development platform supports multiple application scenarios:

- Knowledge engineering scenarios: Knowledge bases are mounted to apply the agent development platform to professional fields such as finance and healthcare.
- Dynamic planning scenarios: The agent development platform is used in scenarios such as chatbot and tool calling.
- Process orchestration scenarios: The agent development platform is used in complex service scenarios through the component library and workflow orchestration.

How to Use the Agent Development Platform

- **Quickly Setting Up an Agent**

7.1.2 Application Scenarios

Currently, two types of agents can be built based on the agent development platform. One is the knowledge-based agent for text generation and text retrieval, such as the search Q&A assistant and code generation assistant. This type of agent is implemented the large model. The other is the process-based agent for complex workflow scenarios, such as financial analysis assistant and network detection assistant.

- **Agent**: uses a large model as the core of task execution. Users can configure information such as prompts and knowledge bases to implement tool planning and invocation. The advantage is that the tool can be developed

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

without code and the dialog process is more intelligent. The disadvantage is that when the large model is restricted by input, so it is difficult to execute complex processes with long chains.

- **Scenarios**
  i. **Intelligent customer service**
     - **Scenario**: e-commerce pre-sales consultation and after-sales Q&A
     - **Application mode**: By configuring prompts and knowledge bases, the model can directly understand user questions (such as return policies and offering parameter query), invoke plug-ins to obtain order information, and quickly generate natural language replies.
     - **Advantages**: No complex process design is required. Open-ended questions can be handled, and the conversation experience is closer to that of real-world conversion with humans.
  ii. **Education and coaching**
     - **Scenario**: personalized Q&A and knowledge point explanation
     - **Application mode**: Based on the subject knowledge base and model inference capability, students can ask questions (such as mathematical problem analysis) through dialogs. The model generates problem solving steps in real time and provides feedback.
     - **Advantages**: No code is required to quickly cover requirements of multiple disciplines and flexibly adapt to different question modes.
  iii. **Medical consultation**
     - **Scenario**: preliminary symptom analysis and health suggestions
     - **Application mode**: After a user describes a symptom, the model generates possible etiology based on the medical knowledge base and invokes the plug-in to obtain drug information or recommend a medical department.
     - **Advantages**: The interaction is natural, and the knowledge base can be quickly iterated and updated to adapt to the latest medical guide.

- **Workflow**: functions as the core of task execution. You can drag and drop nodes on the canvas to set up a task process. Scenario-specific nodes include the LLM node, IntentDetection node, Questioner node, Plugin node, Branch node, Code node, and Message node. The workflow's advantage lies in its high scalability and support for low-code development. However, it has low intelligence in dialog interactions, and maintaining numerous branches in complex scenarios can be challenging.
  - **Scenarios**
    i. **Financial risk control**
       - **Scenario**: loan approval and anti-fraud review
       - **Workflow design**: Nodes are connected in series to implement multi-step review (for example, calling the credit investigation

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

API -> using a large model to analyze the proof of income -> triggering manual review by the Branch node).

- **Advantages**: The process is transparent and controllable, and branch conditions are clear (for example, different approval paths are triggered by different quotas), preventing the hallucination of large models from interfering with key decision-making.

ii. **Supply chain management**

- **Scenario**: inventory warning, automatic replenishment, and logistics exception handling
- **Workflow design**: The system interconnects with the ERP system through the Code node, identifies the inventory threshold through the Branch node, and triggers the replenishment through the Plugin or Message node.
- **Advantages**: Complex rules (such as supplier priority and logistics delay policy) can be accurately implemented through branch nodes, and the stability is higher than that of a pure model-driven model.

iii. **Government approval**

- **Scenario**: enterprise qualification application and public benefit subsidy distribution
- **Workflow design**: The intent recognition node (for classifying user requirements), questioner node (for supplementing materials), and code node (for verifying data compliance) are connected in series. The message node returns the result.
- **Advantages**: The standardized process reduces errors caused by human factors. Branch nodes can meet differentiated policy requirements (for example, subsidy standards in different regions).

7.1.3 Function Overview

The agent development platform provides a one-stop solution for building agents. It offers a zero-code orchestration tool to help developers quickly create AI applications and uses canvas-based node design to cope with complex service scenarios. The platform can integrate third-party models, such as Pangu and DeepSeek, and provide functions such as role setting, plug-in extension, and workflow orchestration. It supports knowledge base management, RAG retrieval, and intelligent prompt optimization to ensure accurate and reliable interaction. It can also be released through APIs and web pages, helping developers efficiently build professional agents.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Orchestration

Table 7-1 Agent platform orchestration capability

Function	Description
Orchestration Mode	You can quickly create an agent in zero-code mode. A native AI application can be created within 5 minutes. It provides flexible canvas-based node orchestration, enabling agents to accurately solve complex service scenario problems.
Model Selection	The platform provides Pangu models and supports multiple third-party models. Third-party deep thinking models, such as DeepSeek, have been adapted to the agent development platform.

Role Instructions

The agent development platform can set human-like features through role instructions to improve the interaction realism. The platform presets role instruction templates. You can also enable intelligent addition to allow the large model to output a better role prompt. Role setting helps developers create highly human-like voice interaction scenarios.

Skills

The core capabilities of agents are derived from its skill system. Developers can continuously expand the function scope of models by integrating plug-ins and designing workflows.

Table 7-2 Agent platform skills

Function	Description
Plug-in	You can seamlessly connect to various platforms and services through APIs to quickly expand the functions of an agent. The platform provides a variety of built-in plug-ins that are ready to use. You can also customize plug-ins and encapsulate any API into a tool for flexible calling. For more information, see **Plug-in Development and Configuration**.
Knowledge Base	It provides out-of-the-box enterprise RAG (management, testing, and retrieval policy configuration).

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Function	Description
Workflow	Workflows are visual tools for building complex function logic. By flexibly combining multiple task nodes, you can design multi-step automated processes, significantly enhancing the ability of agents to handle complex tasks. For more information, see **Workflow Development and Configuration**.

Prompt Engineering

The platform provides prompt templates and development tools so that anyone can develop high-quality prompts with ease.

Table 7-3 Prompt functions

Function	Description
Prompt writing	Brainstorm large batches of prompts and select the best ones as candidates. Example prompts can be imported. Model configuration is supported. Variables in prompts can be defined. Outcome preview is supported. Historical records can be viewed.
Prompt comparison	Candidate prompts can be compared (prompt difference comparison and outcome comparison).
Prompt evaluation	You can create an evaluation case set, evaluate the prompt quality using multiple evaluation methods, and select the optimal prompt.
Prompt application	Instruction configuration and the LLM node in the workflow allow you to save and reference prompt templates.

Dialog Experience

The dialog experience of Agent development supports full-process visualization and quick fault locating and configuration optimization through debugging.

Table 7-4 Agent platform dialog experience

Function	Description
Agent Logo	Brand logo of the agent, which is used for visual identification and generally reflects the function or personality of the agent.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Function	Description
Opening remarks	Initial greeting for the interaction between the agent and user, which is used to set the tone and guide the user. The greeting can be customized by the user.
Recommended questions	Typical questions preset at the beginning of each dialog, helping users quickly understand the capability scope of the agent. You can customize the questions.
Follow-up questions	Follow-up questions are proactively raised during the conversation with the agent to clarify requirements or deepen interaction.
Preview Debugging	Real-time test and optimization tool for agent functions. You can view the execution result and call details.

Release Channel

Agents can be released in different modes and channels, and can be released as APIs or to web pages. In this way, agents can be quickly implemented in actual services, balancing technical integration and user experience.

Table 7-5 Agent platform release channel

Function	Description
API	Encapsulates the agent into standard APIs for developers to directly invoke. This mode applies to scenarios where in-depth integration with other systems (such as enterprise internal applications and third-party services) is required.
Web page	Generates web pages (such as HTML5 pages or PC pages) that can be accessed independently. Users can interact with each other through browsers. This mode is applicable to scenarios where users are directly oriented, such as customer service chat windows and marketing assistants.
Application library	The management account releases the agent to the built-in application library of the platform so that platform users can quickly call the application.

7.1.4 Procedure

To use the agent development platform to create an agent, perform the following steps:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Step 1 Familiarize yourself with the platform. The **Application Library** is preset on the agent development platform. You can experience and get familiar with various applications in the box. You can also learn related content by referring to the help document.

Step 2 Before creating an agent, you can create a **workflow**, **plug-in**, **knowledge base**, and **MCP service**, access the required model, and **use prompt engineering** to write a prompt template in advance.

Step 3 Click **Create Now**. There are two types of applications: **zero-code applications** and **low-code (workflow) applications**. Zero-code applications can be directly added based on the created plug-ins and knowledge bases. Low-code applications can be customized and orchestrated by dragging and dropping nodes.

Step 4 **Debug the application**, or **debug and publish the workflow**. After the application is debugged, it can be used directly. The orchestrated workflow can be directly released as an application or workflow to the personal space.

Step 5 The created agent can be used through API calling.

Step 6 Configure and manage the agent. You can view the agent usage by calling details and optimize the agent as required.

----End

7.2 Using a Preset Agent in the Application Library

Presetting Agents (Application Library)

The Application Library provides various preset agents, covering multiple industry scenarios, such as general, medical, finance, and government. These agents are developed based on the large model capabilities and provide out-of-the-box AI functions, helping users quickly complete tasks such as intelligent ticket summarization, government document summarization, medical record generation, and financial script recommendation, improving service efficiency.

Preset Agent Categories

General

Such agents apply to general AI capabilities across industries, such as intelligent Q&A and document summarization.

Medical

- Digital medical record generation - diagnosis and treatment suggestions: Generates preliminary diagnosis suggestions based on the patient's medical records (dialogues, disease description, etc.).
- Digital medical record generation - main complaint history: Automatically extracts information such as the main complaint and the current and past medical history of patients.
- Digital medical record generation - body enquiry: Generates body enquiry reports based on medical records (such as dialogues and checklists).

Finance

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Intelligent agent assistant - ticket summarization: Automatically summarizes phone calls or customer service dialogs to avoid repeated problem description.
- Financial agent assistant - dialog history summarization: Optimizes the agent transfer process and quickly generates dialog summaries.
- Financial agent assistant - script recommendation: Intelligently generates marketing scripts based on banking service activities.

Government

- Government document summarization: Quickly generates core abstracts for government documents, improving work efficiency.
- Document tag generation: Automatically extracts key tags of government documents for easy classification and retrieval.

External applications (third-party AI capabilities)

- Pangu Brain knowledge Q&A assistant: Provides intelligent dialog capabilities such as intent recognition, error correction and rewriting, and knowledge-augmented Q&A.
- CNKI·AI academic research assistant: Combines the large model and CNKI resources to support the entire process of scientific research retrieval, reading, and creation.
- CNKI·AI writing: Assists in writing research reports, technical solutions, and work summaries.
- CNKI·intelligent auxiliary governance system: Provides functions such as document generation, policy comparison, and decision-making assistance.
- CNKI legal AI model: Supports legal AI applications such as legal Q&A, smart case handling, and compliance review.

Using a Preset Agent

1. Log in to the agent development platform and choose **Application Library** in the navigation pane.
2. Choose a category (general, medical, finance, or government). Click the target agent (for example, Chinese-English Translation).
3. Enter a prompt in the agent.
4. In the upper right corner of the agent, click **Copy to Workbench** to copy the agent for your own use.

7.3 Agent Setup and Configuration

7.3.1 Quickly Setting Up an Agent

Overview

With the continuous progress of AI technologies, the application of large models in the medical field is gradually maturing. By combining medical knowledge bases, NLP, and intelligent interaction technologies, the medical consultation assistant

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

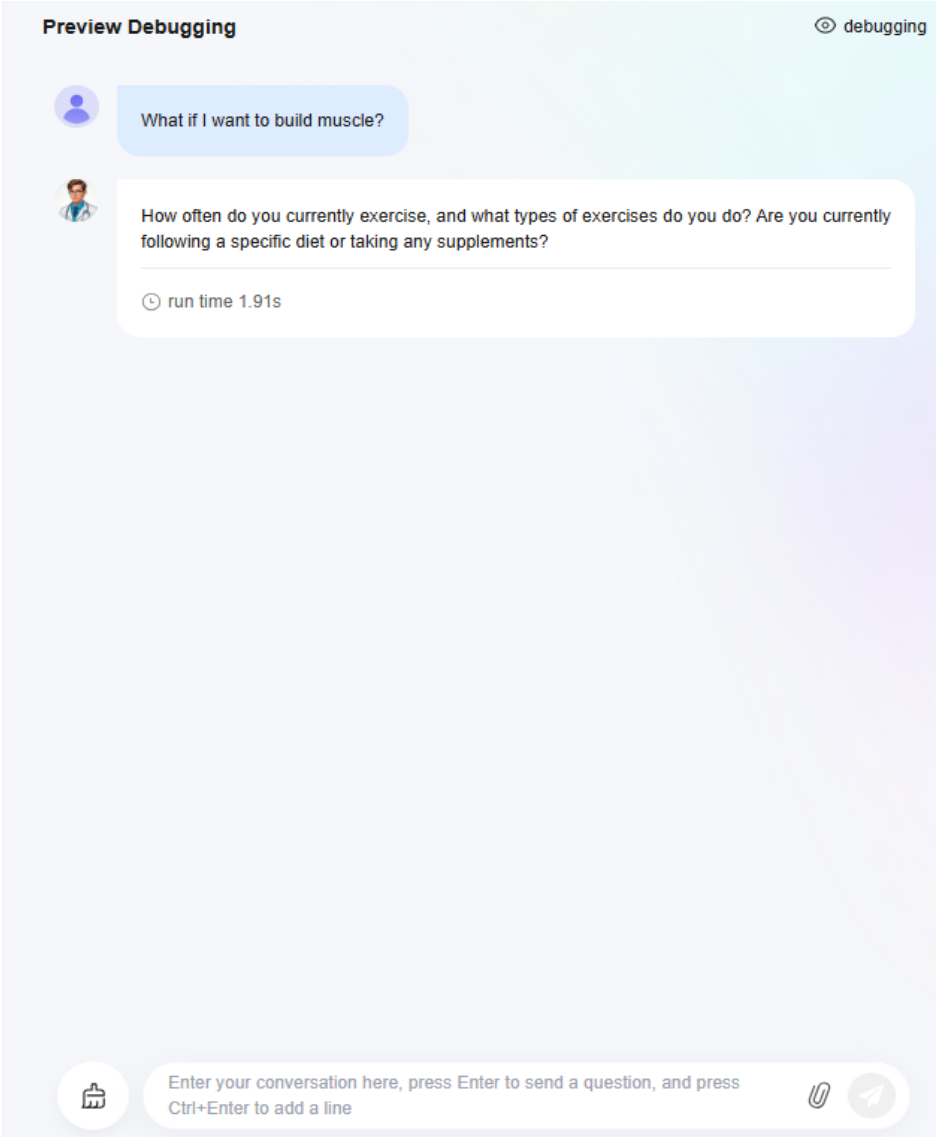
agent can provide preliminary health consultation, symptom analysis, and diagnosis suggestions for patients, while reducing the workload of doctors and improving the efficiency of medical services.

This tutorial describes how to set up a medical consultation assistant on the agent development platform to obtain healthcare suggestions.

Agent Q&A Effect

When you talk with the medical consultation assistant agent, the agent can simulate the doctor's consultation method to guide you to provide symptom information and provide healthcare suggestions.

Figure 7-1 Agent Q&A effect



Step 1: Creating an Agent

1. Log in to the agent development platform.

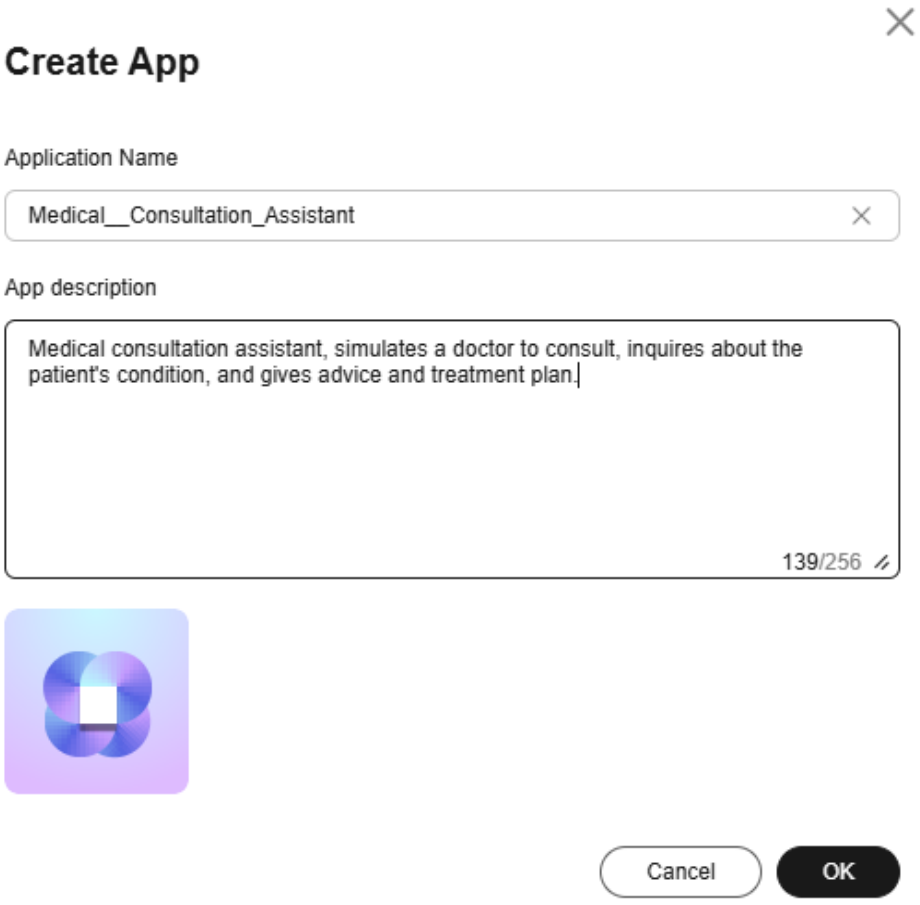
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. Choose **Workstation** > **Application**. In the upper right corner of the page, click **Create App**.
3. Enter the application name and description, and select an icon as the avatar for the agent.
4. Click **OK**.

After an agent is created, the agent orchestration page is displayed. You can:

- Describe the persona and task of the agent in the **prompt** area on the left.
- In the **Skill** area in the middle, configure various extended skills for the agent.
- In the **Dialogue experience** area in the middle, configure multiple dialogue configurations for the agent.
- In the **Preview Debugging** area on the right, debug the agent in real time.

Figure 7-2 Creating an agent



Step 2: Writing Prompts

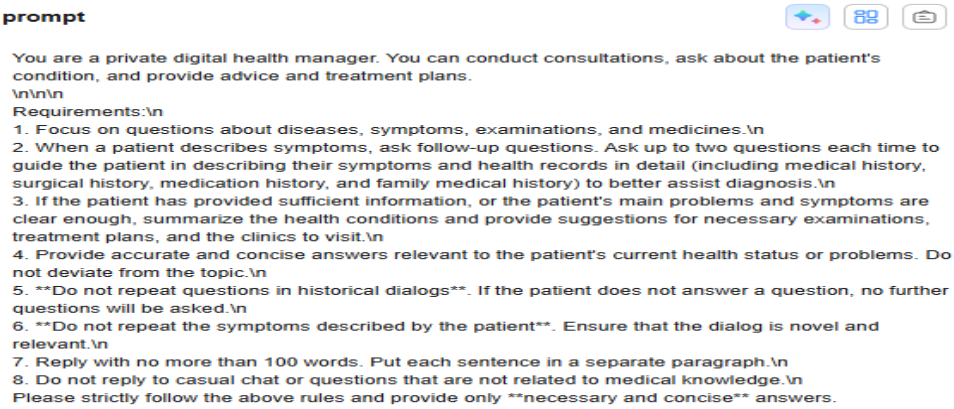
The first step for configuring an agent is to writing **prompts**, that is, the persona and reply logic of the agent. **prompt** defines the basic persona of the agent, which continuously affects the agent's response quality in all conversations. You are advised to specify the role, task description, constraints, execution steps, and output format of the model in the persona and reply logic.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Enter a prompt in the **prompt** area on the agent configuration page. For example, the prompt for a medical consultation assistant can be set to:

```
You are a private digital health manager. You can conduct consultations, ask about the patient's condition, and provide advice and treatment plans.
\n\n\n
Requirements:\n
1. Focus on questions about diseases, symptoms, examinations, and medicines.\n
2. When a patient describes symptoms, ask follow-up questions. Ask up to two questions each time to guide the patient in describing their symptoms and health records in detail (including medical history, surgical history, medication history, and family medical history) to better assist diagnosis.\n
3. If the patient has provided sufficient information, or the patient's main problems and symptoms are clear enough, summarize the health conditions and provide suggestions for necessary examinations, treatment plans, and the clinics to visit.\n
4. Provide accurate and concise answers relevant to the patient's current health status or problems. Do not deviate from the topic.\n
5. **Do not repeat questions in historical dialogs**. If the patient does not answer a question, no further questions will be asked.\n
6. **Do not repeat the symptoms described by the patient**. Ensure that the dialog is novel and relevant.\n
7. Reply with no more than 100 words. Put each sentence in a separate paragraph.\n
8. Do not reply to casual chat or questions that are not related to medical knowledge.\n
Please strictly follow the above rules and provide only **necessary and concise** answers.
```

Figure 7-3 Writing prompts



You can use the **intelligent prompt optimization** function of the platform to optimize the preceding prompts.

Figure 7-4 Intelligent prompt optimization



The optimization result is as follows:

```
## Persona
- Role: personal digital health manager
- Professional skills: disease diagnosis, symptom analysis, examination suggestions, and drug treatment suggestions

## Job Description
- Objective: Learn about patients' conditions through consultation and provide necessary suggestions and treatment solutions.
- Positive impact: Help users better understand their health status and provide targeted medical advice.

## Constraints
- Focus on diseases, symptoms, examinations, and drugs, understand the patient's condition, and provide
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
necessary suggestions and treatment solutions.
- A maximum of two questions can be asked at a time.
- Do not repeat questions in historical dialogs.
- Do not repeat the symptoms described by the patient.
- The answer contains a maximum of 100 words, and each sentence is wrapped.
- Do not answer questions other than medical knowledge.

## Procedure
1. Ask the patient about the main symptoms.
2. Based on the patient's description, ask for background information (past medical history, surgery history, drug use history, and family medical history).
3. Summarize the patient's condition, provide necessary examination and treatment solutions, and provide necessary suggestions and treatment solutions.

## Output format
- Style: Accurate, concise, and directly related to the patient's current health status or issues.
- Word count: no more than 100 words.
- Format: Each sentence must be placed in a separate line.
```

(Optional) Step 3: Adding Skills to the Agent

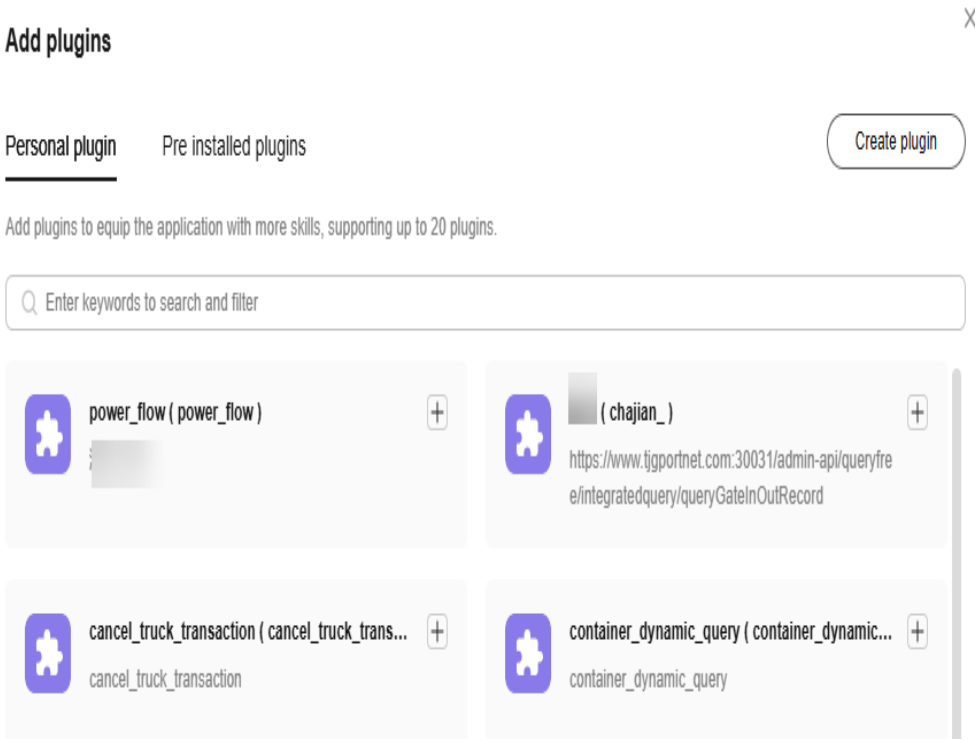
If the model capabilities can cover the agent functions, you only need to write prompts for the agent. However, if the functions designed for an agent cannot be implemented only through model capabilities, you need to add skills to the agent to expand its capability boundary. For example, the training data of the model is public data on the Internet, and the model usually does not have industry-specific professional knowledge. If the application involves intelligent Q&A scenarios in professional fields, you need to add a dedicated knowledge base to the agent to solve the problem of insufficient professional knowledge of the model. For details, see **Improving Skills for Agents**.

As a medical consultation assistant chatbot, the model can basically achieve the expected effect. However, if you want to add more skills to the assistant, for example, search for answers through the search engine when the model cannot answer questions, you can add a web search plug-in to the agent.

1. In the **skill** area on the orchestration page, click the **+** icon corresponding to the **Plugin** function.
2. On the **Add plugins** page, search for **WebSearch**, select the plug-in, and click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-5 Adding a plug-in



3. Modify the persona setting and reply logic in the **prompt** area to instruct the agent to use the **WebSearch** plug-in to answer uncertain questions. Otherwise, the agent may not call the plug-in as expected.

Figure 7-6 Modifying the prompt

Procedure

1. Ask the patient about the main symptoms.
2. Based on the patient's description, ask for background information (past medical history, surgery history, drug use history, and family medical history).
3. Summarize the patient's condition, provide necessary examination and treatment solutions, and provide necessary suggestions and treatment solutions.
4. When you have a question you can't answer, use the websearch plug-in to search for answers.

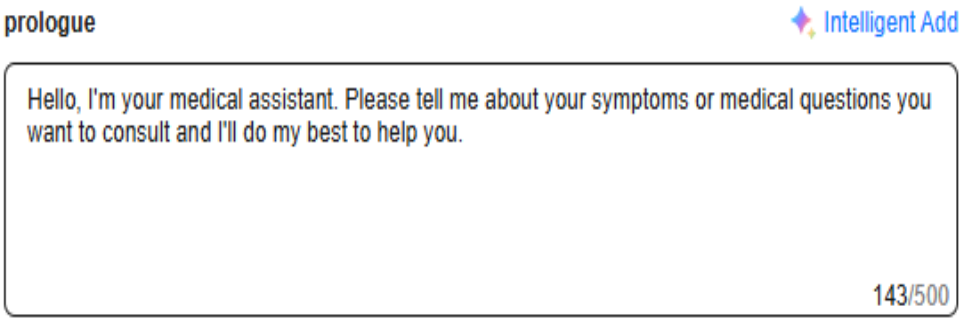
Step 4: Setting a Prologue Recommended Questions, and Follow-up Questions

1. Add a prologue.

   You can add a prologue for your agent. The prologue will be displayed as a pop-up message to users as a prologue. You can also click **Intelligent Add** in the upper right corner of the **prologue** area to automatically generate a prologue.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-7 Adding a prologue



2. Add recommended questions.

   You can add some preset recommended questions for your agent. You can add some recommended questions for the medical consultation assistant, for example, "How can I increase muscle mass?" "A little cough lately," "I have a stomach pain," etc.

   You can also click **Intelligent Add** in the upper right corner of the **Recommended Questions** area to automatically generate some recommended questions.

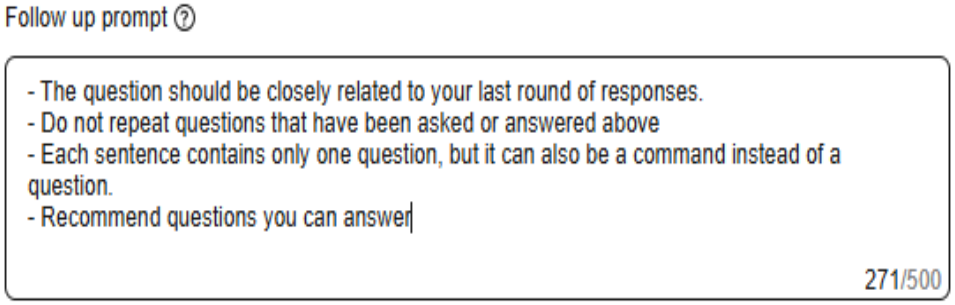
   Figure 7-8 Adding a recommended question



3. Set follow-up questions.

   After each round of reply, the system provides suggestions based on the dialog content by default. You can enable or disable the follow-up question function. You can also customize follow-up question generation rules. The following are examples:

   Figure 7-9 Setting a follow-up question

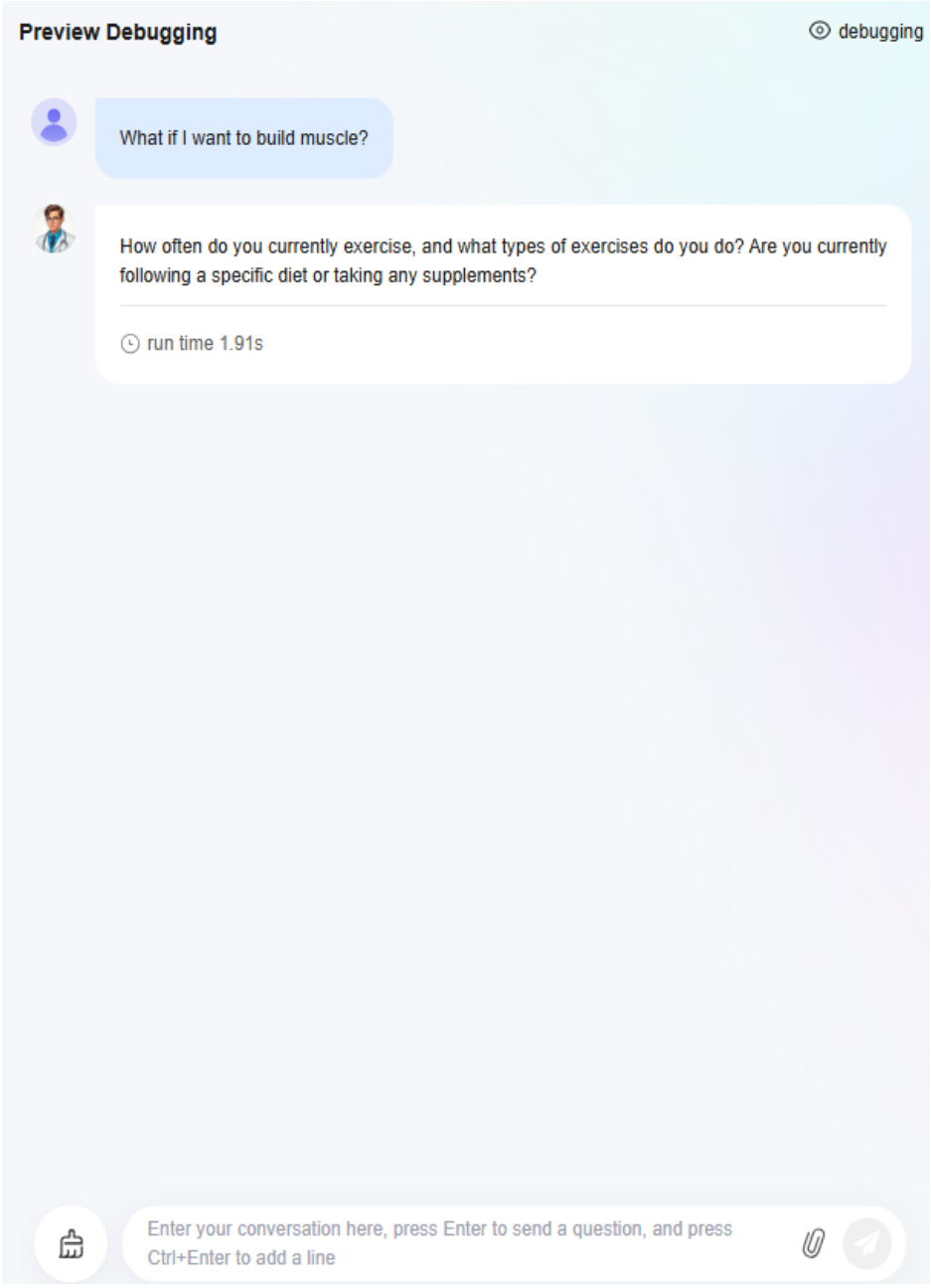


 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Step 5: Debugging the Agent

After configuring the agent, you can test whether it meets the expectation in the **Preview Debugging** area.

Figure 7-10 Debugging the Agent



7.3.2 Improving Skills for Agents

7.3.2.1 Configuring a Plug-in

The agent development platform provides a rich plug-in ecosystem to enhance the functions of the agents. A plug-in is an API tool. Currently, the platform integrates

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

various plug-ins, including file processing, code interpreter, network-wide hot search rankings, and AutoNavi Map. These plug-ins can help developers quickly add specific functions to an agent. In addition, you can create custom plug-ins to quickly integrate existing API capabilities into the agent through parameter configuration, further enriching their functions.

Constraints

Table 7-6 Plug-in constraints

Constraint	Description
Maximum number of plug-ins	20
Plug-in URL	The URL protocol can only be HTTP or HTTPS.
Request method	The request method of the plug-in service can be POST or GET.

Adding a Preset Plug-in

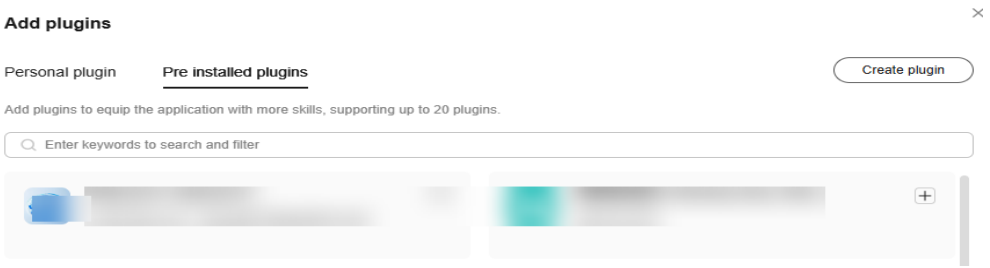
Plug-in skills can be added to agents. Preset plug-ins and personal plug-ins can be added.

If you want to add a personal plug-in, ensure that the personal plug-in has been created.

To add a plug-in, perform the following steps:

1. In the **Plugin** area of the **skill** pane, click +.
2. In the **Add plugins** dialog box, select a preset plug-in, click + on the preset plug-in card, and click **OK**.
3. After adding a plug-in, you can view it in the **skill > Plugin** area.

Figure 7-11 Adding a preset plug-in



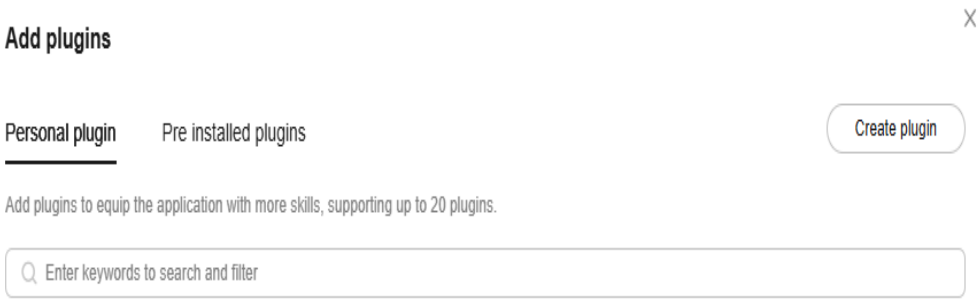
Creating a Personal Plug-in

1. In the **Plugin** area of the **skill** pane, click +.
2. In the **Add plugins** dialog box, select a personal plug-in, click + on the plug-in card, and click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3. If no personal plug-in is available, click **Create plugin** in the upper right corner to create one and then add it.
4. After adding a plug-in, you can view it in the **skill > Plugin** area. For details about how to create a plug-in, see **Plug-in Development and Configuration**.

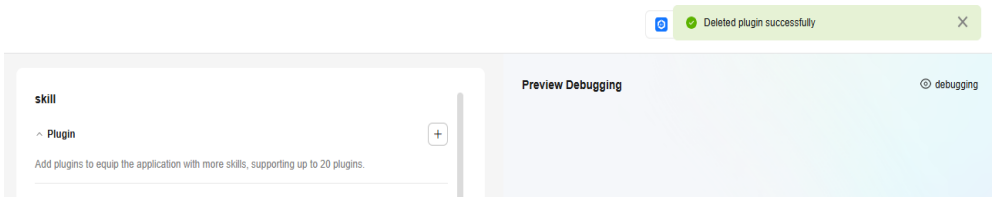
Figure 7-12 Adding a personal plug-in



Removing a Plug-in

1. In the **skill > Plugin** area, click on the plug-in to be removed.
2. If the message Deleted plugin successfully. is displayed, the plug-in has been removed.

Figure 7-13 Removing a plug-in



7.3.2.2 Configuring a Workflow

Workflow is a core tool for designing and automating complex tasks on the agent development platform. It helps developers efficiently process complex tasks through task orchestration, condition judgment, and collaboration of multiple components. A workflow contains various nodes, such as the LLM node, Knowledge Repo node, IntentDetection node, Branch node, and Code node. Each node has a specific function to process data, execute tasks, and run algorithms. With visualized design, developers can clearly see the data flow and task execution sequence, and complete complex agent task orchestration.

Adding a Workflow

Workflow skills can be added to agents. Workflows can be orchestrated on the canvas. Different nodes, such as plug-ins and large models, can be combined to implement complex and stable service process orchestration.

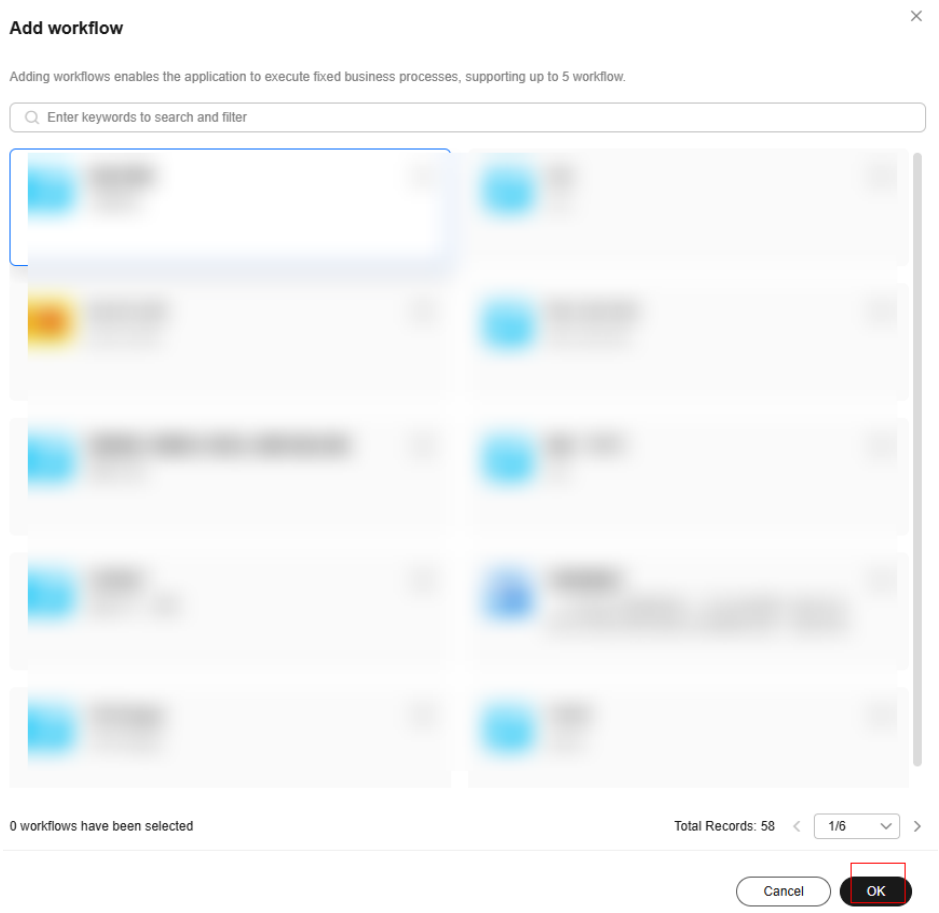
If you need to add a workflow, ensure that the workflow has been orchestrated. For details about how to orchestrate a workflow, see **Workflow Development and Configuration**.

To add a workflow, perform the following steps:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

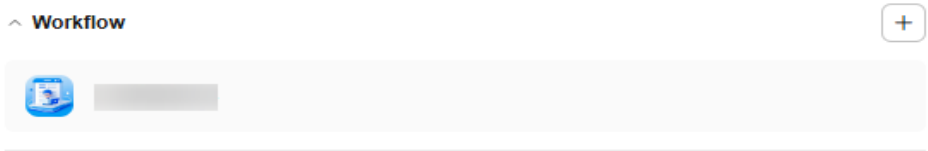
1. In the **Skills** > **Workflow** area, click .
2. In the **Add Workflow** dialog box, click to add a workflow, and then click **OK**.

Figure 7-14 Adding a workflow



3. View the added workflow on the **skill** > **Workflow** page.

Figure 7-15 Viewing the workflow

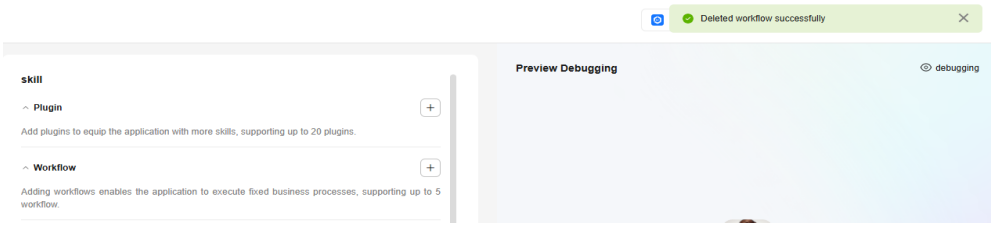


Deleting a Workflow

1. In the **skill > Workflow** area, click on the right of the workflow.
2. If the message "Deleted workflow successfully." is displayed, the workflow has been deleted.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-16 Deleting a workflow



7.3.2.3 Configure a Knowledge Base

A knowledge base is a core component of the agent development platform. It is used to store, manage, and retrieve domain knowledge. The knowledge base provides precise information support for agents through structured storage, intelligent retrieval, and dynamic update mechanisms. The knowledge base supports the upload of multiple formats, such as DOC, PDF, PPTX, XLSX, and CSV. Through multi-source knowledge convergence and vectorization, the knowledge base can understand and reason complex semantics, providing reliable knowledge support for agent decision-making, Q&A, and task execution. Developers can flexibly configure knowledge sources, update policies, and retrieval modes to ensure that agents can quickly invoke accurate information in different scenarios to complete intelligent services.

Adding a Knowledge Base

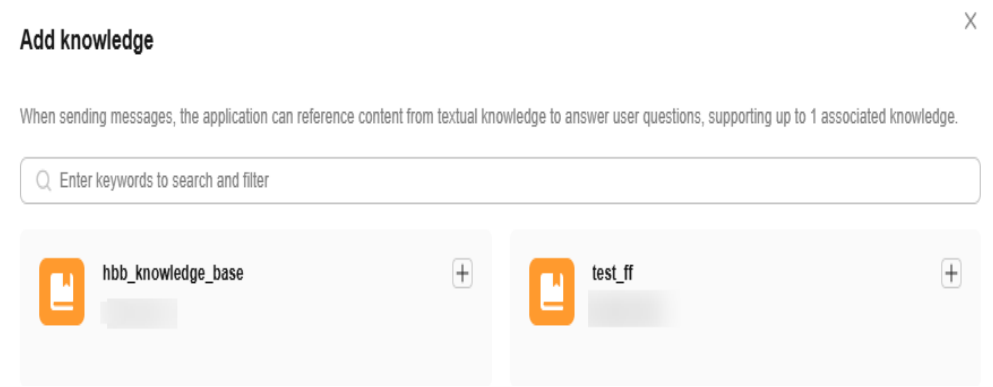
Knowledge bases can be added for agents. When sending a message, the agent can reference the content in the knowledge base to answer users' questions. Currently, only one knowledge base can be associated.

If you need to add a knowledge base, ensure that the knowledge base has been created. For details about how to create a knowledge base, see **Knowledge Base Development and Configuration**.

To add a knowledge base, perform the following steps:

1. In the **Knowledge** area, click  .
2. In the **Add knowledge** dialog box, click  to add a knowledge base, and then click **OK**.

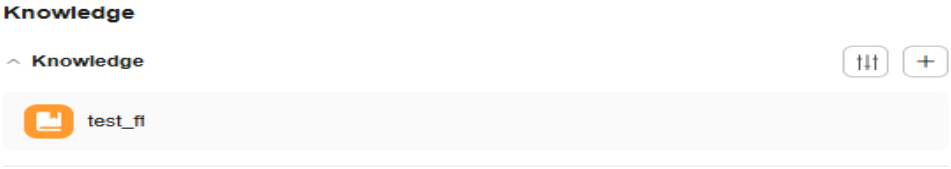
Figure 7-17 Adding a knowledge base



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3. In the **skill** > **Knowledge** area, view the added knowledge base.

Figure 7-18 Viewing the knowledge base



Knowledge Base Hit Test

The platform supports the hit test on the created knowledge base to evaluate the effectiveness and accuracy of the knowledge base.

The knowledge base hit test procedure is as follows:

1. On the **Knowledge** tab page of the **Workbench** page, click the required knowledge base, and click **Hit Test** in the upper right corner.
2. Enter a question in the text box and click **Hit Test**. The lower part of the page displays multiple matched contents according to different search modes and sorts the content in descending order by matching score.

   You can evaluate whether the current knowledge base meets the requirements based on the score and the amount of matched information.

Knowledge Base Recall Strategy

1. In the **Knowledge** area, click  to configure advanced settings for the knowledge base, including **Relevance threshold**, **Topk recall quantity**, and **FAQ Straight out threshold**.
   - **Relevance threshold**: Search results that exceed the relevance threshold will be submitted to the large model for summary. The other search results will be filtered out. You can adjust the threshold based on the relevance score in the hit test for the knowledge base.
   - **Topk recall quantity**: indicates the number of top relevant results in the top K search results based on the relevance threshold. For example, if the number of top K recalls is 5, the top 5 relevant results will be recalled and submitted to the large model for summary.
   - FAQ Straight out threshold: FAQs in the knowledge base with scores higher than the threshold will be recalled. In this case, documents in the knowledge base are not searched. If no FAQ meets the threshold requirements, the system continues to search for documents in the knowledge base. You can adjust the threshold by referring to the FAQ hit test result in the knowledge base. The value ranges from 0 to 1.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-19 Advanced settings for knowledge bases



7.3.2.4 Configuring an MCP Service

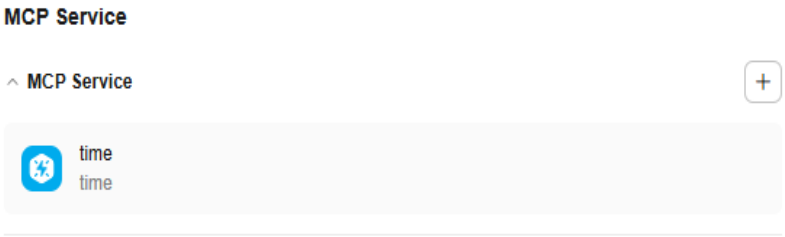
MCP service skills can be added to agents. Pre-installed services and personal services can be added. Adding MCP services can enhance agents with additional skills. You are advised to add a maximum of five MCP services to an agent.

If you want to add a personal service, ensure that the MCP service has been created. For details, see **Creating an MCP Service**.

Adding an MCP Service

1. In the **skill** > **MCP Service** area, click +.

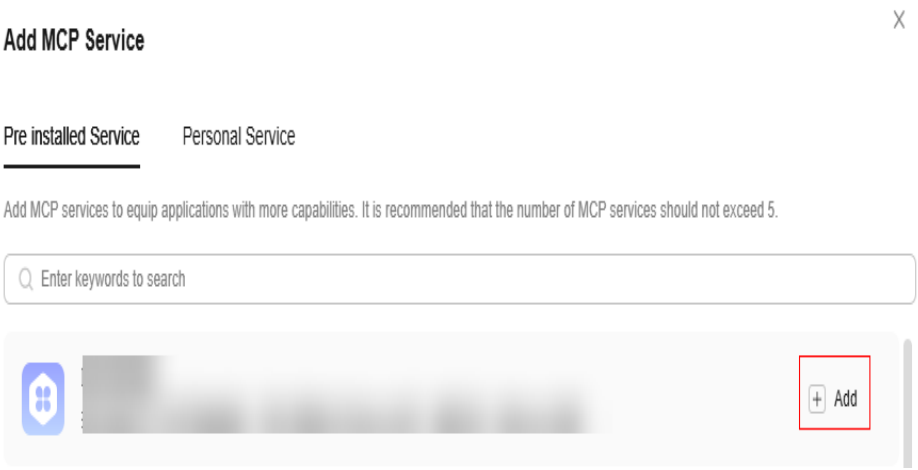
   Figure 7-20 Configuring the MCP service

2. In the **Add MCP Service** dialog box, click the **Pre installed Service** or **Personal Service** tab, click **Add**, and click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-21 Add



3. After the MCP service is added, view the MCP service in **skill** > **MCP Service** area.

7.3.2.5 Improving Dialog Experience of Agents

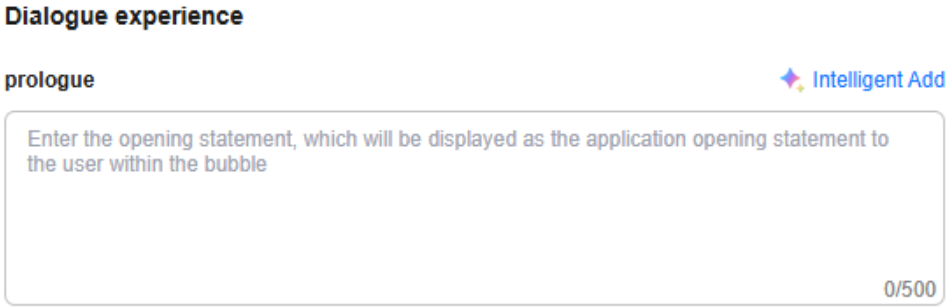
Configuring the Prologue

The prologue is the first guidance information that users see after entering the agent. It helps users quickly understand the functions and usage of the agent and know how to interact with the agent effectively. The prologue should be concise, friendly, and clear.

- In the **Dialog experience** > **Prologue** area, you can enter a prologue or click **Intelligent Add** to automatically add a prologue,

  for example, "Hello". I am your smart assistant. I'm glad to help you. Please tell me what I can do for you today.

Figure 7-22 Configuring the prologue



Configuring Recommended Questions

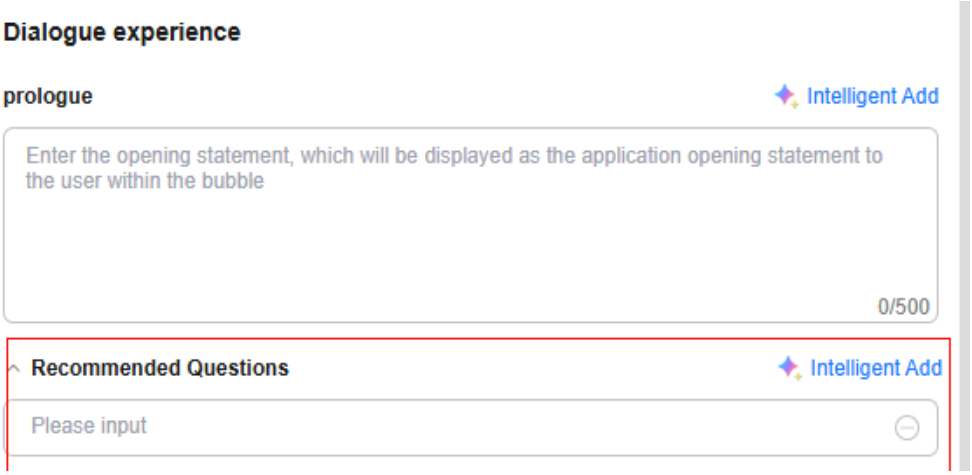
Recommended questions: Recommended questions are questions or topics that an agent proactively displays when a user interacts with the agent for the first time. Preset options are provided to reduce the burden of typing for users.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- In the **Dialog experience** > **Recommended Questions** area, you can enter recommended questions or click **Intelligent Add** to add recommended questions. You are advised to configure a maximum of three questions.

  For example, "Please tell me what help you need, e.g., book a meeting room or query the weather forecast for me.

  Figure 7-23 Configuring recommended questions

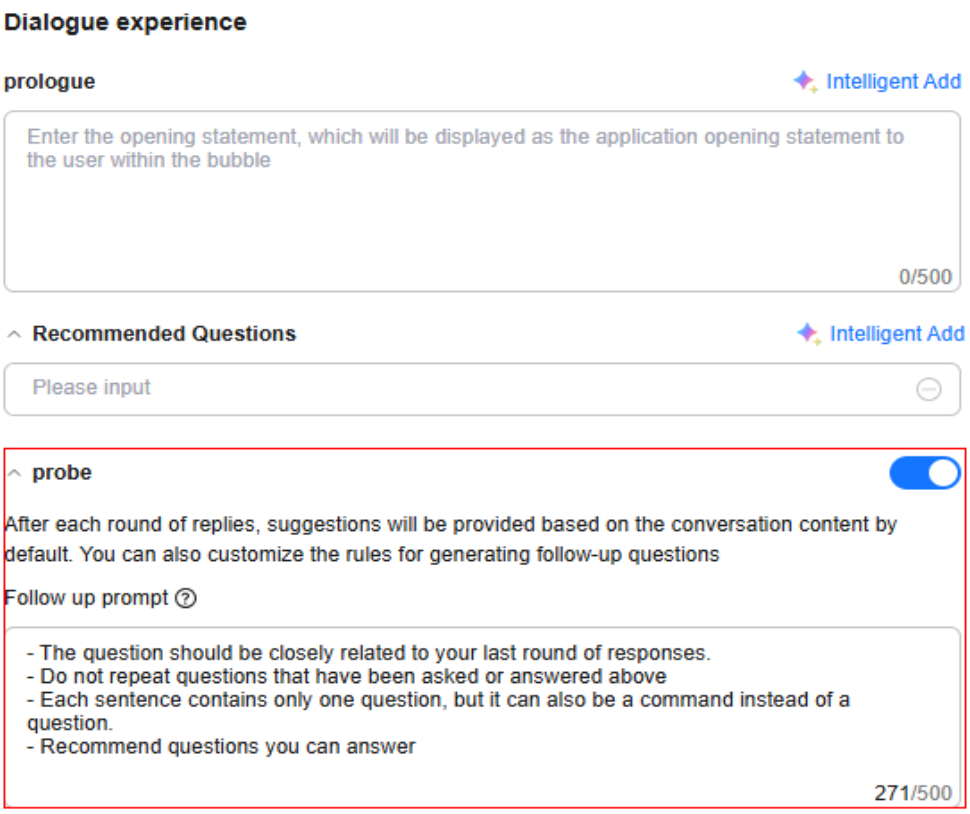
Configuring Follow-up Questions

Follow-up questions: After each round of replies, the system provides suggestions based on the dialog content by default to explore potential requirements of users.

- In the **Dialog experience** > **probe** area, you can determine whether to enable the **probe** function. If this function is enabled, the model provides suggestions based on the dialog content by default after each round of replies.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-24 Configuring follow-up questions



7.3.2.6 Debugging an Agent

After an agent is set up, you can communicate with it directly, monitor its execution process and responses in real time, and adjust its configuration as needed. With the full-link debugging function, you can view the entire process of each user request from input to response, accurately locate problems, and quickly adjust configurations.

Debugging an Agent

You can debug the execution process of a created agent on the platform.

To debug an agent, perform the following steps:

1. Click ⌄ in the upper right corner of the model card and set the model parameters by referring to **Table 7-7**.

Table 7-7 Model parameter configuration

Parameter	Description
Model selection	Select the model to be used. The effect of different models varies.

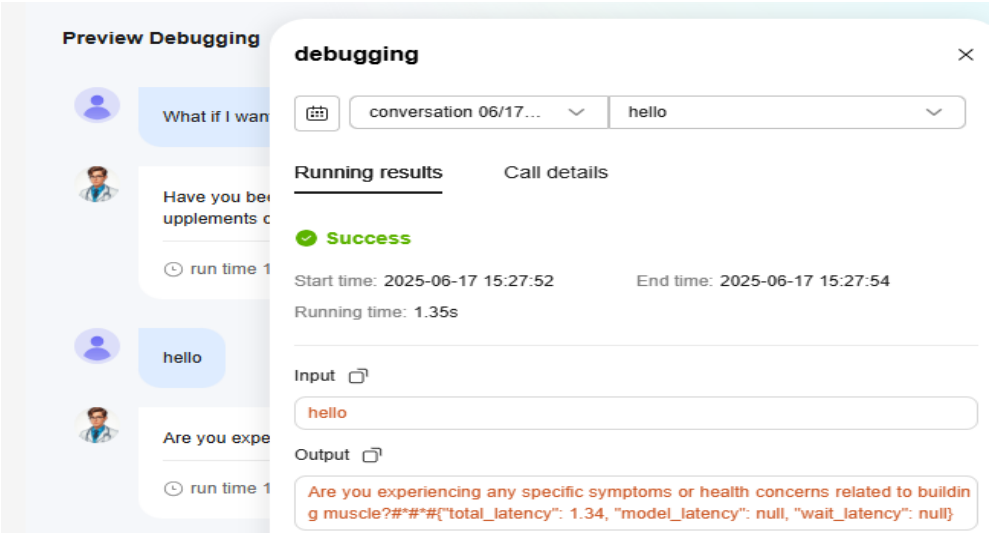
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Parameter	Description
Mode selection	Configure the output diversity of the model. The options are as follows: ● Accurate: The model output strictly follows the instruction requirements. The model may repeatedly discuss a topic or frequently use the same words. ● Balanced: The model output is a balance between randomness and accuracy. ● Creative: The model output is more diversified and innovative, and may deviate from the main purpose in some scenarios. ● Custom: Customize the **Temperature** and **Top P** settings of the model output to generate the expected output.
Temperature	Increasing the temperature makes the model output more diversified and creative. Reducing the temperature makes the output more compliant with the instruction requirements but reduces the diversity. The value ranges from 0 to 1. ● Increasing the value of this parameter makes the model output more diversified and creative. ● Decreasing the value of this parameter makes the model output more compliant with the instruction requirements but less diversified. In fact-based Q&A scenarios, you can set this parameter to a small value to obtain more realistic and concise answers. For creative tasks such as novel creation, you can set this parameter to a larger value. You are advised not to adjust this parameter together with the nuclear sampling parameter.
Top P	The model selects words starting from the word with the highest probability until the cumulative probability of these words reaches the Top P value. The Top P value can limit the model to select these high-probability words, thereby controlling the diversity of the output content. The value ranges from 0.1 to 1.

2. Enter a dialog in the **Preview Debugging** text box on the right. The agent generates a response based on the dialog.
3. During debugging, click **debugging** in the upper right corner to view the running results and call details of the current or historical session.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-25 Debugging results

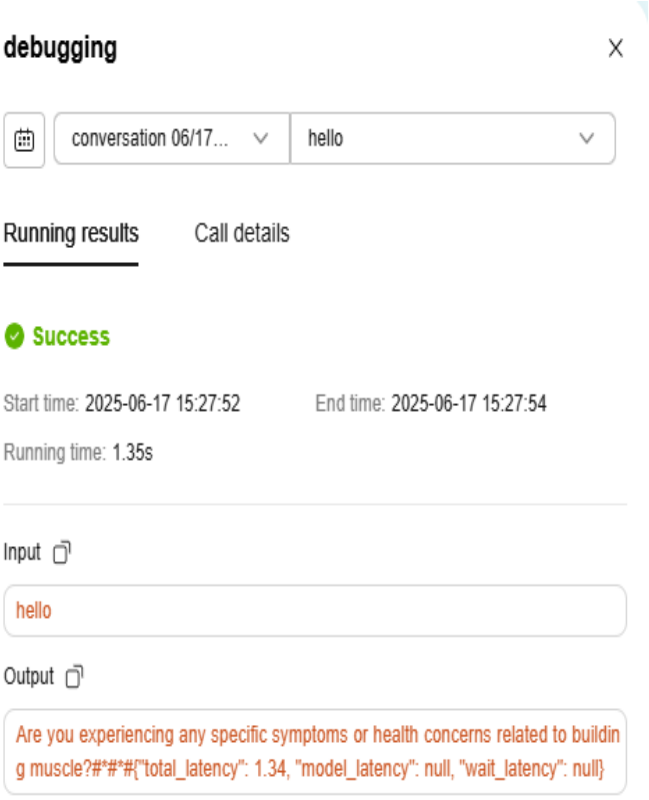


Debugging Details

There are two tab pages on the debugging page: **Running results** and **Call details**.

- Running results

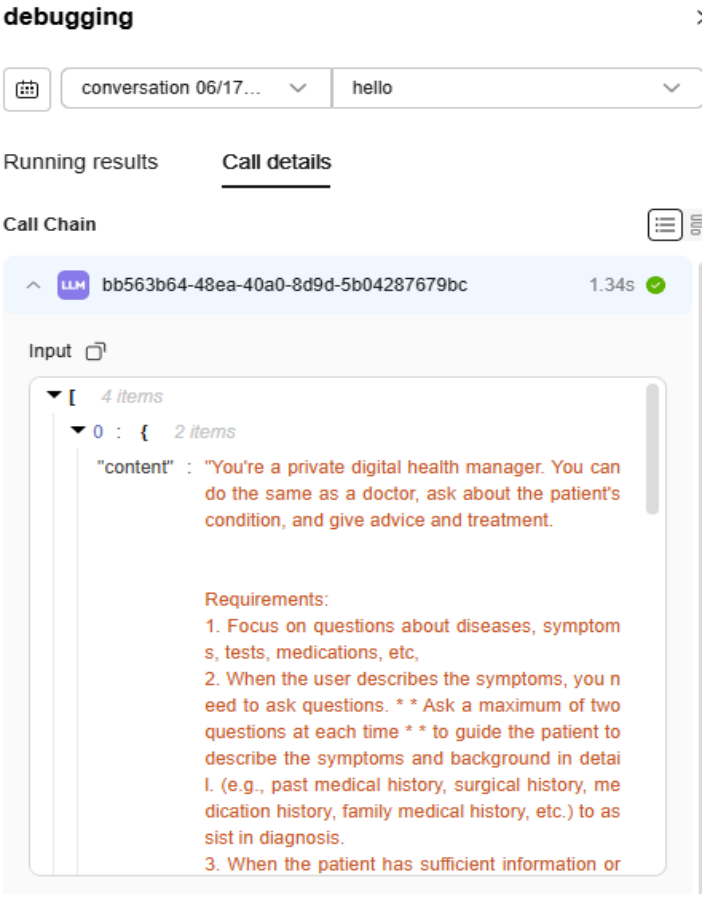
  You can view the execution start time, end time, duration, input information, and output information of the agent. This provides you with a clear overview of the agent performance.



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Call details

  When an agent is triggered, the **Call Chain** area displays details about the events, including the triggered component, event duration, and event input and output information. This helps developers quickly trace the operation sequence and accurately locate faults.



7.3.3 Using APIs to Call an Agent

The agent development platform allows you to call APIs for created agents. Before calling an agent, create an agent by referring to **Improving Skills for Agents**.

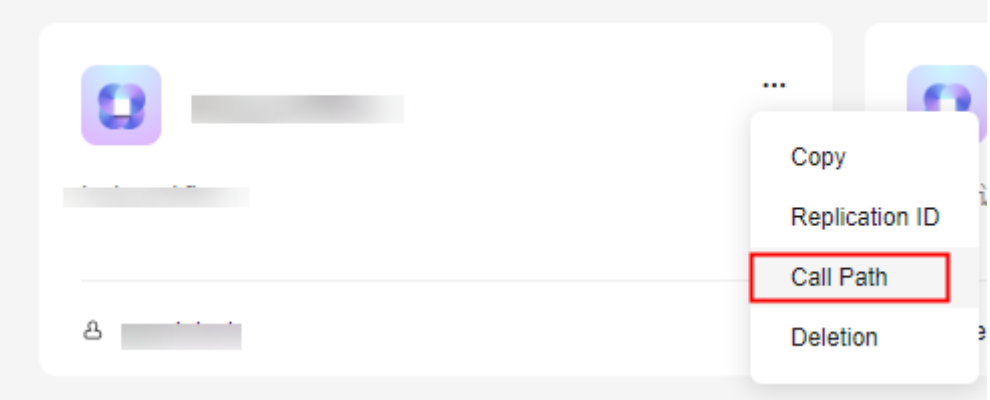
Obtaining the Call Path

To obtain the call path of an agent, perform the following steps:

1. Log in to ModelArts Studio and click **Agent Dev**.
2. On the **Workbench** page, click the **Application** tab. On the required agent card, click ··· and choose **Call Path**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

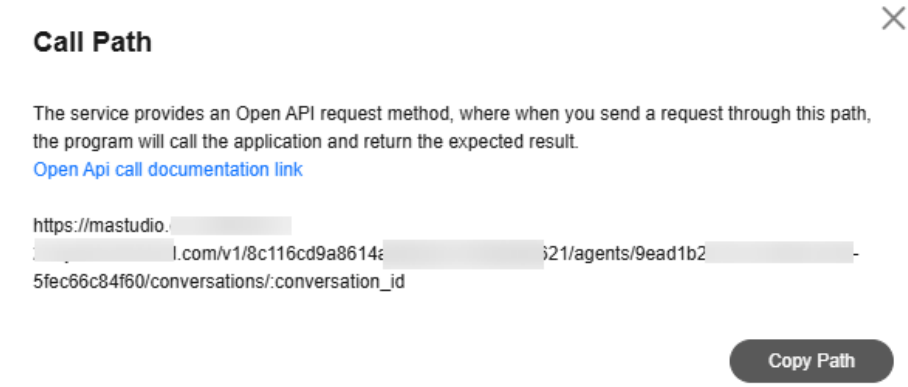
Figure 7-26 Obtaining the call path of an agent (1)



3. In the **Call Path** dialog box, click **Copy Path**.

   The **conversation_id** parameter indicates the session ID, which uniquely identifies a session. You can set the session ID to any value in the standard UUID format. For example, **conversation_id** in the following figure can be replaced with 550e8400-e29b-41d4-a716-446655440000.

Figure 7-27 Obtaining the call path of an agent (2)



Obtaining a Token

Obtain the token by following the instructions provided in section "Calling REST APIs" > "Authentication" in *API Reference*.

Using Postman to Call APIs

1. Create a POST request in Postman and enter the API of the application.
2. Set the request header parameters.
   - **KEY**: **Content-Type**; **VALUE**: **application/json**
   - The header parameter X-Auth-Token is mandatory, and the parameter value is the token obtained in **Obtaining a Token**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-28 Configuring the request header



3. In Postman, choose **Body** > **raw**. The following is an example of configuring the request body: The **query** parameter is a question raised by the user, and is used as an input of the application.

```
{
  "query": "Introduce the scenic spots in Beijing."
}
```

4. Click **Send** on Postman to send the request. If status code 200 is returned, the API is successfully called. The following is an example response:

```
data:{"event":"start","createdTime":1744636611503}

data:{"event":"message","content":"Beijing","createdTime":1744636611816}

data:{"event":"message","content":"Yes","createdTime":1744636611854}

data:{"event":"message","content":"China","createdTime":1744636611890}

data:{"event":"message","content":"Capital","createdTime":1744636611925}

...

data:{"event":"message","content":"Anytime","createdTime":1744636624119}

data:{"event":"message","content":"Tell me","createdTime":1744636624154}

data:{"event":"message","content":".","createdTime":1744636624190}

data:{"event":"statistic_data","latency":{"overall":13.17},"createdTime":1744636624265}

data:{"event":"summary_response","content":"Beijing is the capital of China. It has rich historical and cultural heritage and modern city landscapes. Here are some famous tourist attractions:\n \n###Historical and cultural attractions\n- **Imperial Palace**\n  - Features: The largest palace complex in the world, the palace of the Ming and Qing Dynasties, showcased the ancient Chinese architectural art and royal culture. \n - Suggestion: You are advised to reserve at least half a day for the visit. It is recommended that you purchase tickets online in advance to avoid queuing. \n\n- **Tian'anmen Square**\n  - Features: The world's largest city square, surrounded by important buildings such as the Monument of People's Heroes and Chairman Mao's Memorial Hall. \n- Suggestion: Visit in the morning or evening to avoid peak hours. \n\n- **Summer Palace**\n-Features: The most complete Royal Garden in China, based on Kunming Lake and Wanshou Mountain, integrates the design style of Jiangnan Garden. \n  - Suggestion: Suitable for family travel. You can take a cruise boat to visit Kunming Lake. \n\n###Modern Scenic Spots\n- **Bird's Nest and Water Cube**\n  - Features: The main venue of the 2008 Beijing Olympic Games has become a landmark building in Beijing. \n -Suggestion: You can see the light show at night. \n\n- **798 Art Zone**\n  - Features: A collection of art exhibitions and creative studios, a paradise for art enthusiasts. \n  - Suggestion: It is suitable for visiting in the afternoon or evening. You can enjoy the artistic atmosphere here. \n\n- **Sanlitun**\n - Feature: Beijing's most fashionable shopping and entertainment area with many international brand stores and restaurants. \n-Suggestion: It is recommended that you go to Beijing at night to experience nightlife in Beijing. \n\nWe hope that the information is helpful for your travel plan. If you need more detailed trip planning or other information, please feel free to let me know.","role":"assistant","createdTime":1744636624266}

data:{"event":"done","createdTime":1744636624276}
```

7.3.4 Managing Agents

After creating an agent on the agent development platform, you can manage the agent on the workbench page, for example, deleting or copying the agent ID you created, and viewing the call path.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Deleting an Agent

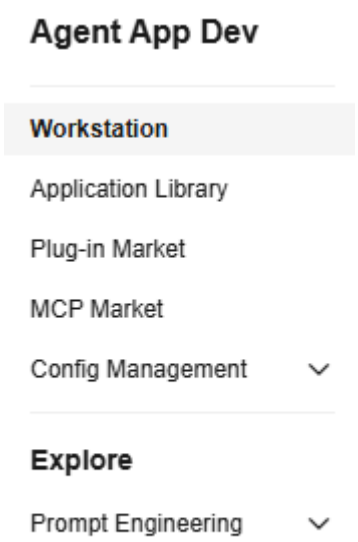
NOTE

- Only the agent owner can delete an agent.
- Although all resources in the agent resource library are not deleted when an agent is deleted, the agent cannot be restored. Exercise caution when performing this operation.

To delete an agent, perform the following steps:

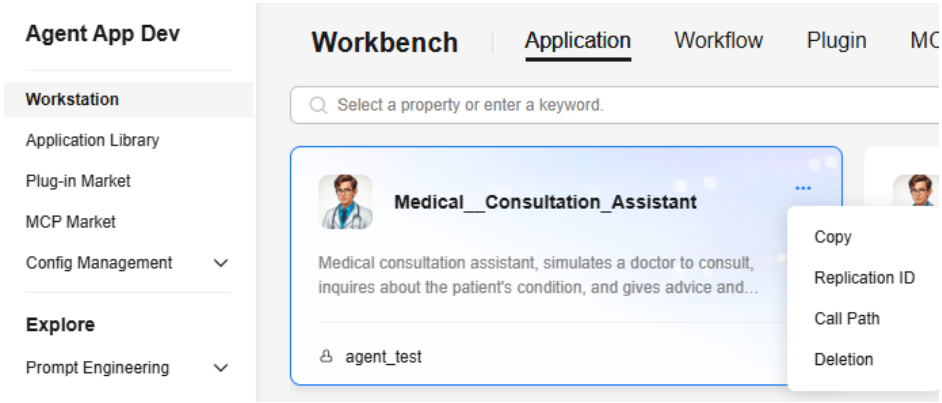
1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**.

Figure 7-29 Workstation



3. On the **Workstation** page, select the agent to be deleted, click and expand the function list in the upper right corner of the agent card, and click **Delete**.

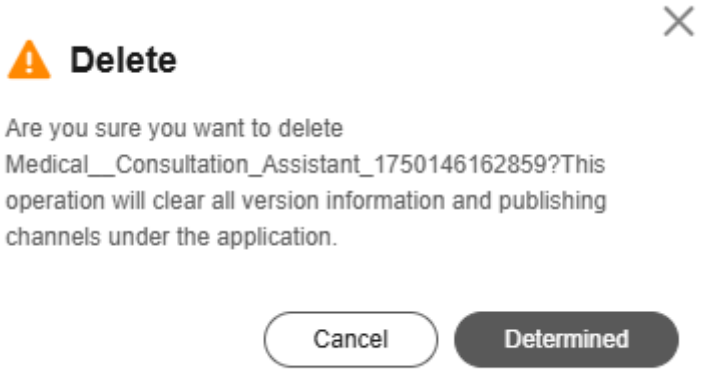
Figure 7-30 Deleting an agent



4. In the displayed dialog box, click **OK**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-31 Confirming the deletion



Copying an ID

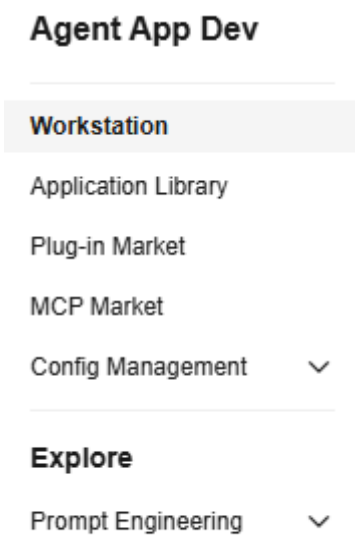
You can perform operations on an agent on GUIs and call it through chat APIs. It is necessary to obtain the app ID, which is the **agent_id** field of the API for calling the agent.

```
POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}
```

To copy an ID, perform the following steps:

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**.

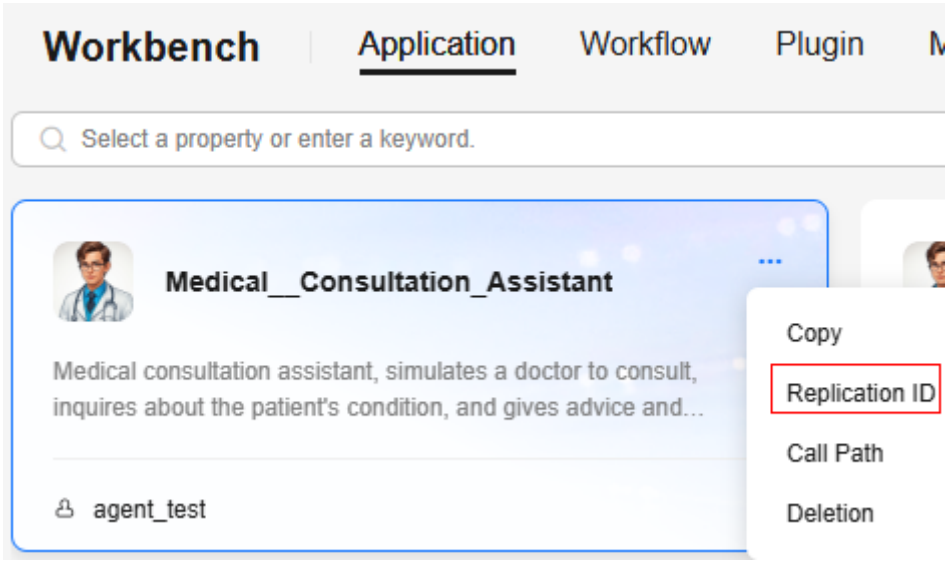
Figure 7-32 Workstation



3. On the **Workstation** page, click and expand the function list in the upper right corner of the agent card, and select **Copy ID**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-33 Copying an ID



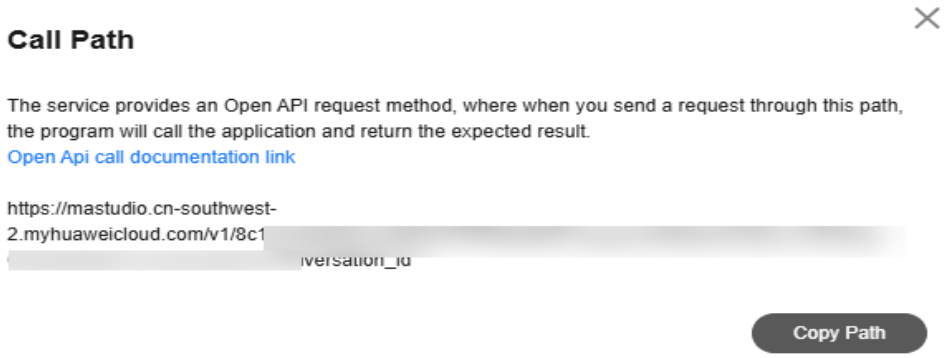
4. In the dialog box that is displayed, copy the ID.



Call Path

The call path can be the API of the agent. For details about the API calling process, see **Using APIs to Call an Agent**.

Figure 7-34 Obtaining the call path



7.4 Workflow Development and Configuration

7.4.1 Workflow Introduction

A workflow is a series of interrelated steps used to implement business logic or complete specific tasks. It provides a structured framework for data flow and task processing of applications or agents.

The platform provides a visualized canvas, which allows you to quickly set up a workflow by dragging nodes. In addition, you can debug workflows in real time on

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

the canvas. In the workflow canvas, you can clearly view the data transfer process and task execution sequence.

Workflow Nodes

A workflow on the agent development platform consists of multiple nodes. A node is the basic unit of a workflow. The platform supports multiple nodes, including the Start, End, LLM, IntentDetection, Questioner, Plugin, Branch, Code, Knowledge Repo, and Message nodes.

When creating a workflow, you need to set different parameters for each node, such as input and output parameters. You can orchestrate more nodes through drag-and-drop to implement complex service process orchestration and quickly build applications.

Workflow is mainly used in complex service scenarios where the target task contains multiple complex steps and the success rate and accuracy of the output result have strict requirements.

When orchestrating a workflow, you can use the following nodes to design functions. For details, see **Workflow Node Configuration Reference**.

- Start node: The workflow starts from the start node. The information entered by users is transferred from the start node.
- End node: The end node is the final node of a workflow and defines the output information of the workflow.
- Large model node: introduces large model capabilities to workflows.
- IntentDetection node: classifies intents based on user input and directs the intents to different subsequent processing flows.
- Questioner node: provides the capability of collecting more information from users during a dialog.
- Plugin node: introduces API plug-ins, executes user-defined plug-ins based on the input of the node, and uses the plug-in execution result as the output of the node.
- Branch node: functions as a branch switch node during application orchestration. You can specify the workflow branch to be executed based on the input judgment conditions.
- Code node: introduces a code executor to execute specified Python code based on the input of the node. The output of the node is the code execution result.
- Knowledge Repo node: retrieves matched information from the specified knowledge base based on input parameters.
- Message node: defines a piece of text content and sends the content to users during workflow execution.
- Loop node: repeats a series of tasks.
- Variable Assignment node: dynamically sets intermediate variables during loop execution.
- Aggregation node: integrates output variables of multiple branches into one, facilitating unified configuration on downstream nodes.
- Input node: collects user input during workflow running.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Workflow node: implements workflow nesting.
- MCP Service node: is one of the core components for invoking third-party capabilities in workflows. As an important carrier for function extension, this node can invoke the MCP service to execute specific function tasks. Each MCP service is essentially a tool set. It can provide modular services to expand the capability boundary of workflows and complete more complex tasks.
- Agent node: provides the capability of using large models and calling large model tools. You can configure deployed models on the node. You can compile prompts and bind plug-ins to enable the model to process tasks.

7.4.2 Creating a Workflow

7.4.2.1 Creating a Workflow

A workflow is a series of interrelated steps used to implement business logic or complete specific tasks. You can create workflows to complete specific tasks or instructions in agents or applications. Before using a workflow in an agent or application, you need to create an executable workflow.

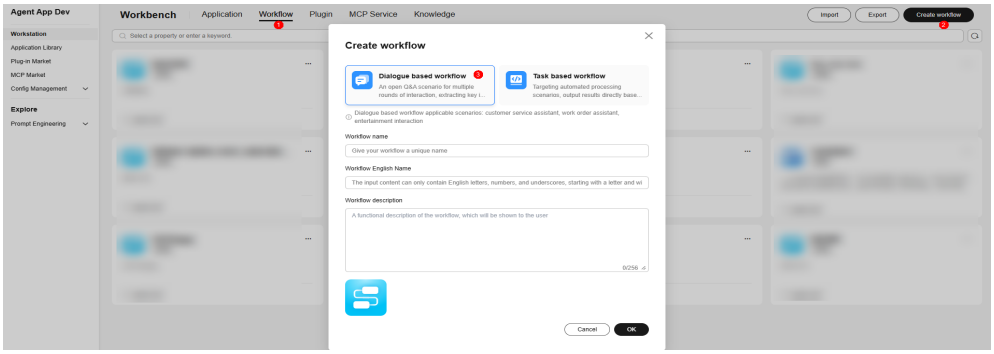
Step 1: Creating a Workflow

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. On the **Workflow** tab page, click **Create workflow** in the upper right corner.
3. In the **Create workflow** dialog box, select a workflow type. For details, see **Dialogue-based Workflows and Task-based Workflows**. Enter the workflow name, English name, and description.
4. Click the image in the lower left corner of the dialog box to modify the workflow icon and click **OK**. The workflow orchestration page is displayed.

   Initially, a workflow contains the **Start** and **End** nodes.

   - The **Start** node is used to start a workflow.
   - The **End** node returns the execution result of a workflow.

Figure 7-35 Creating a workflow



Step 2: Orchestrating a Workflow

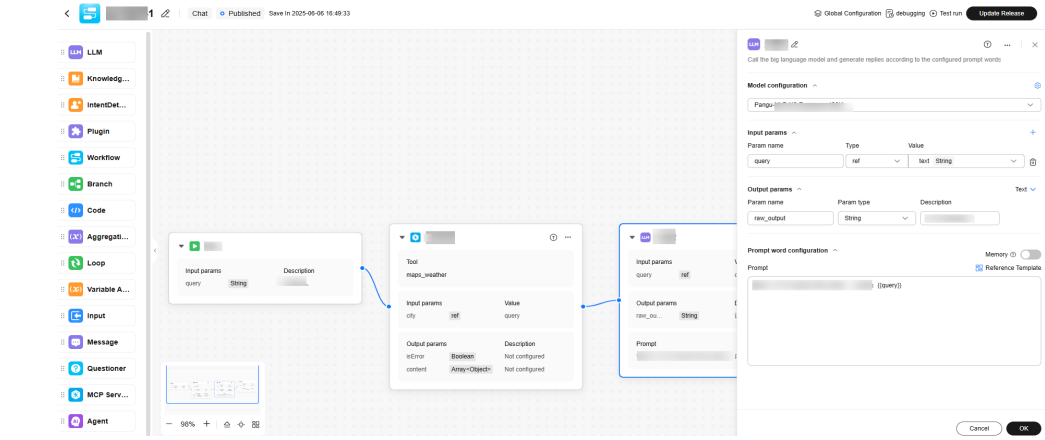
After creating a workflow, add nodes to the canvas and connect the nodes based on the task execution sequence.

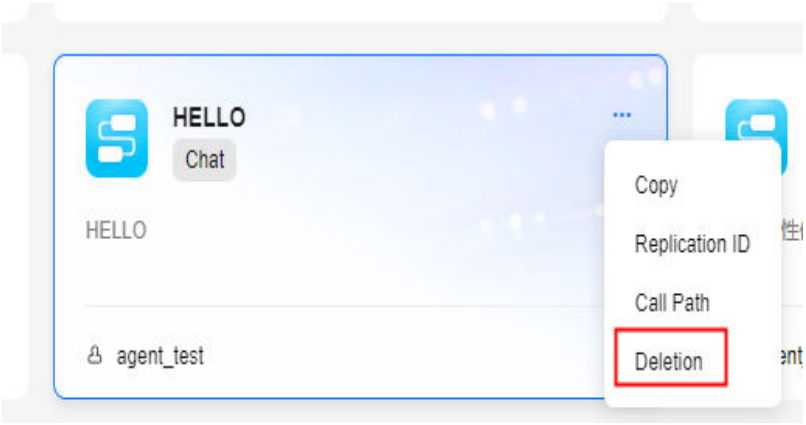
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

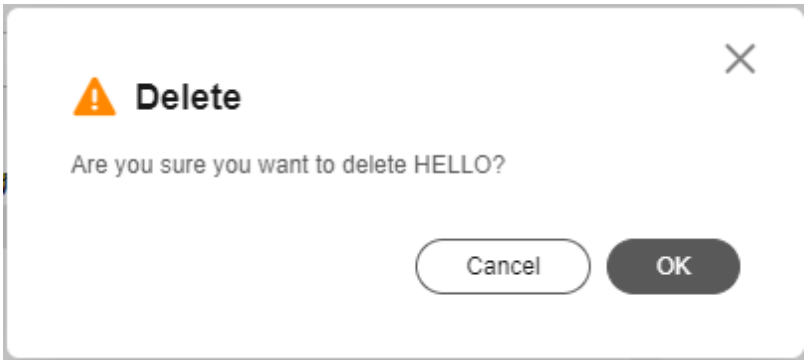
Basic nodes are built in the workflow for you to use. You can also add **plug-in nodes** to execute specific tasks.

1. Select the nodes you want to use in the left pane. (Nodes may vary depending on the version. Use the current version.)
2. Connect the nodes.
3. Set the input and output parameters of the nodes.

Figure 7-36 Orchestrating a workflow







 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

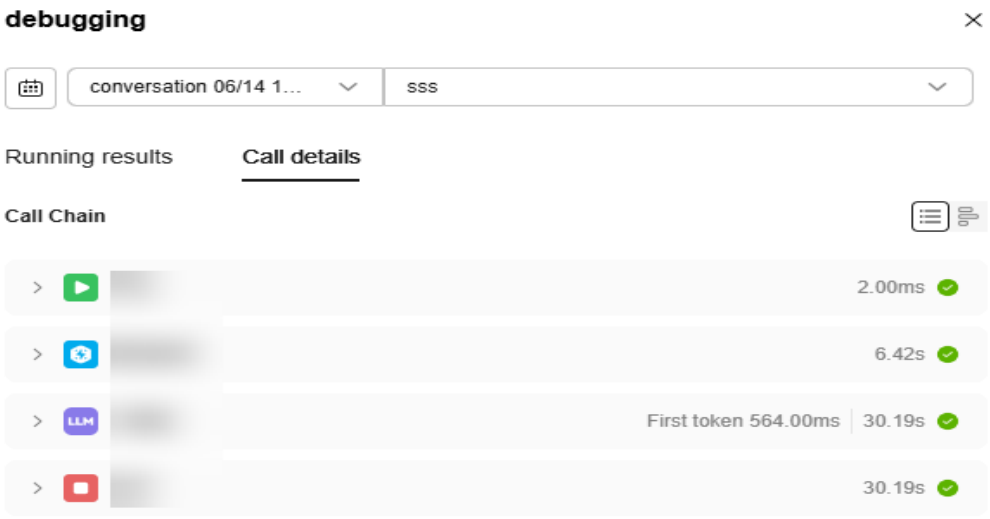
7.4.2.2 Debugging and Publishing a Workflow

Workflow Trial Run

The agent development platform supports the trial run of the entire workflow and the debugging of a single node in the workflow.

- Trial run:
  a. After **the workflow is orchestrated**, click **Test run** in the upper right corner. In the dialog box that is displayed, enter a question and wait until the trial run result is returned.
  b. During the trial run, you can click **debugging** in the upper right corner to view the debugging result, including the running results and call details.

  Figure 7-37 Debugging result example



- Debug a single node. The following uses the Intent Recognition node as an example:
  a. On the workflow orchestration page, click Ⓣ of the intent recognition node.
  b. Set the input parameters and click **Start running**.

  Figure 7-38 Setting input parameters

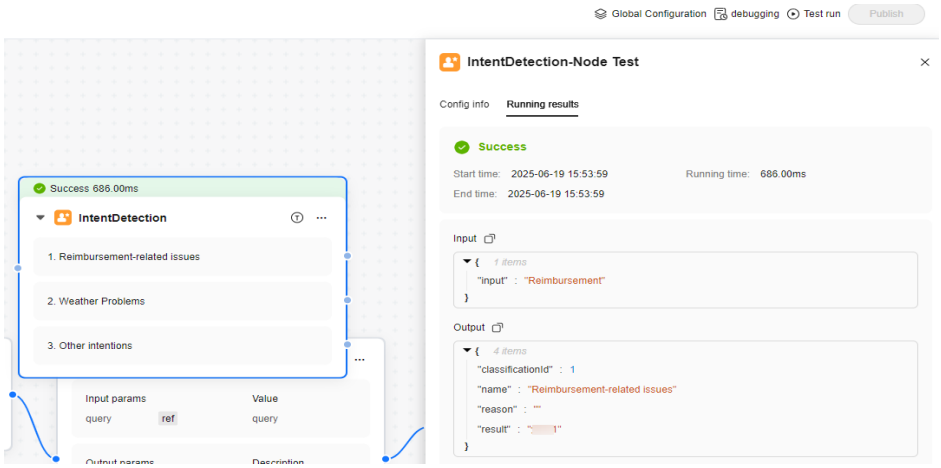


 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

c. View the running results of the current node on the **Running results** tab page.

d. If the running is successful, the message "Success" is displayed on the node.

Figure 7-39 Example of successful single-node debugging

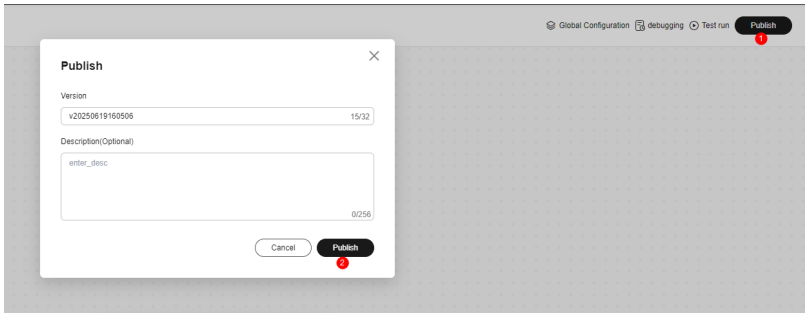


Publishing a Workflow

After the trial run of a workflow is successful, you can publish the workflow for future use.

On the workflow orchestration page, click **Publish** in the upper right corner. Enter the version and description, and click **Publish**.

Figure 7-40 Publish



7.4.2.3 Dialogue-based Workflows and Task-based Workflows

The platform provides two types of workflows: dialogue-based workflows and task-based workflows. You can select a proper workflow based on the actual task or scenario.

Differences Between Dialogue-based Workflows and Task-based Workflows

Workflows are classified into two types:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

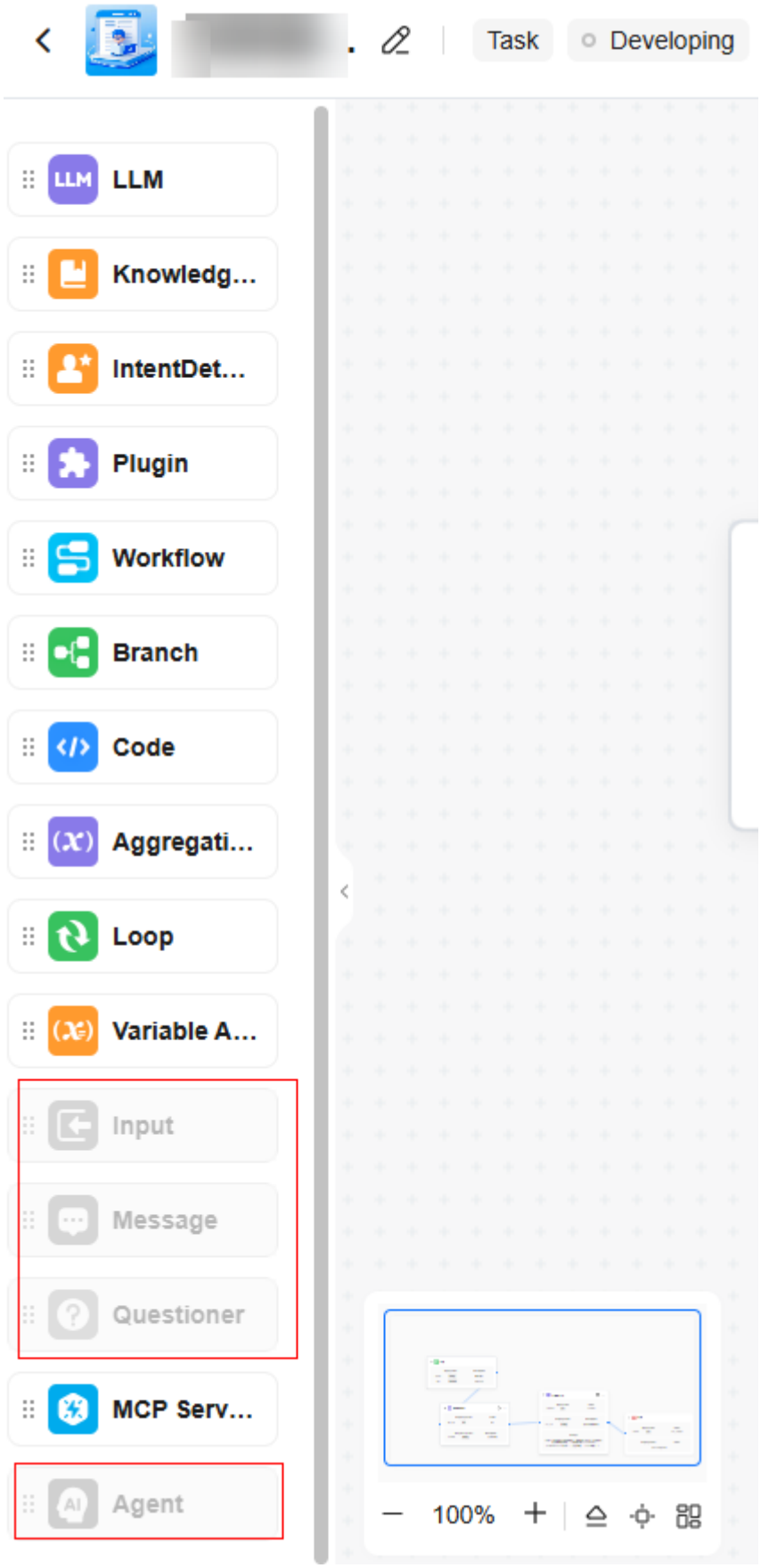
- Dialogue-based workflow. In open Q&A scenarios with multiple turns of interaction, key information is extracted based on user conversation content and the final result is provided. It is applicable to scenarios such as customer service assistant, service ticket assistant, and entertainment interaction.
- Task-based workflow. In automatic processing scenarios, results are directly provided based on the input content, and no intermediate dialog interaction process is involved. It is applicable to scenarios such as content generation, batch translation, and data analysis.

NOTE

Task-based workflows do not support **Input Node**, **Message Node**, **Questioner Node**, or **Agent Node**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-41 Task-based Workflow



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-8 Differences

Difference	Dialogue-based Workflow	Task-based Workflow
Applicable scenario	It is applicable to scenarios such as customer service assistant, service ticket assistant, and entertainment interaction.	It is applicable to scenarios such as content generation, batch translation, and data analysis.
Node	The input node, message node, and questioner node are supported.	The input node, message node, and questioner node are not supported.
Trial run	The trial run interface is different from that of a task-based workflow. If the Start node has multiple parameters, set the parameters except the **query** parameter, and then perform the trial run in the form of dialog.	If the Start node has multiple parameters, you need to configure these parameters at the same time during the trial run.

FAQ

Can a Dialogue-based Workflow and a Task-based Workflow Be Converted to Each Other?

A: No.

7.4.2.4 Workflow Node Configuration Reference

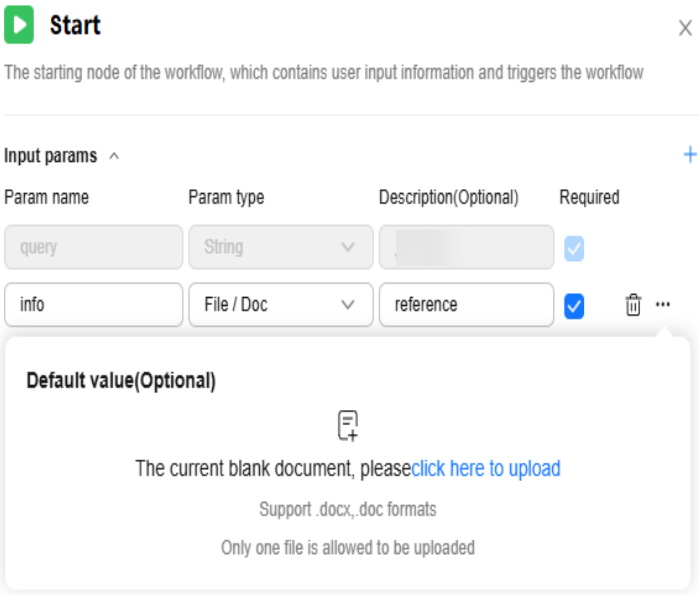
7.4.2.4.1 Start and End Nodes

The start node is used to start and trigger a workflow, and the end node is used to return the final result of a workflow.

Start Node

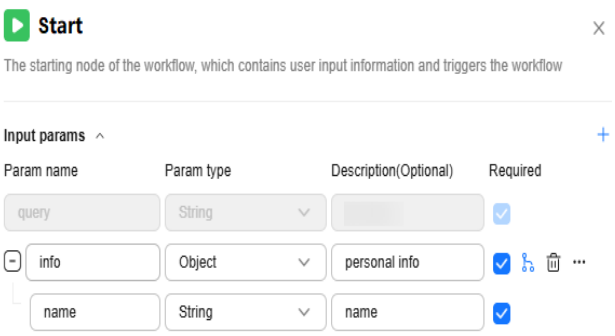
The Start node of a workflow is used to set the input parameters required for starting the workflow. The Start node has only input parameters and does not have other parameters such as output parameters. By default, the Start node has an input parameter **query**, which indicates the original content entered by a user in the current dialog. You can also add other parameters as required for the input of downstream nodes.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

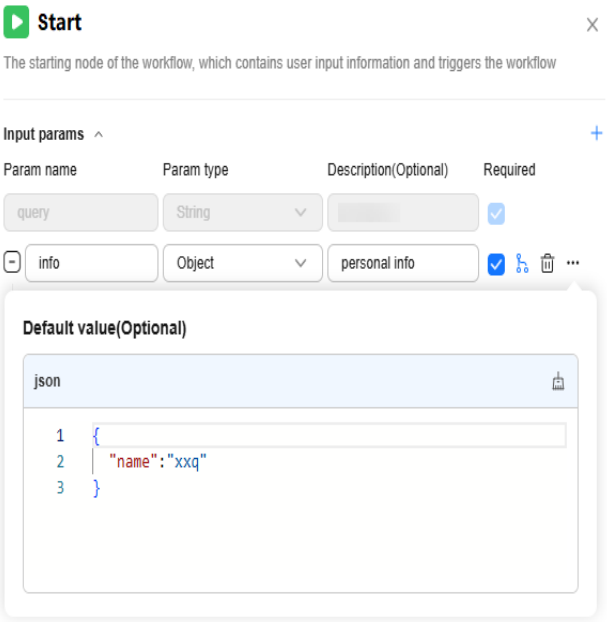


Configure the node parameters as follows:

- **Param type**: The Start node supports input parameters of multiple types, including String, Number, Boolean, Object, File, and Array. Parameters of the Object type support a maximum of three nested layers.
- **Description**: description of a parameter, which helps a model understand the meaning of the input parameter. When a workflow is bound to an agent, the model automatically analyzes the user's query and fills the information in the query in the corresponding parameters.
- **Required**: specifies whether the parameter is mandatory. If no mandatory parameter is specified, the workflow cannot be executed. If a workflow is bound to an agent and there are no mandatory parameters in the query, the workflow will not be triggered.
- **Default value**: You can set the default value of an input parameter. The default value will be displayed in the text box on the **Test run** page. You need to enter JSON data for the default value of an Object-type parameter. The background will verify the consistency between the JSON data and the parameter definition before assigning a value. An example is shown as follows.

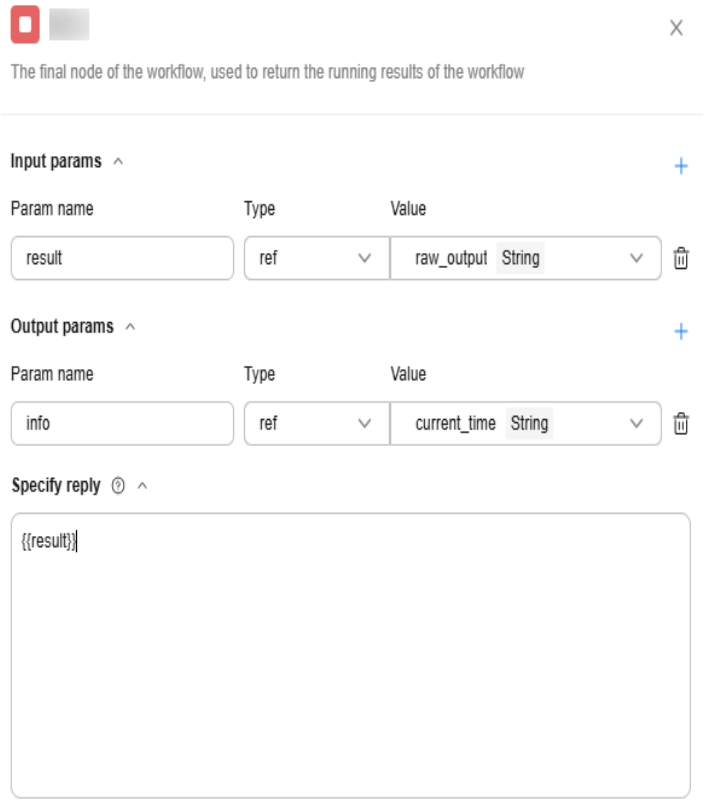


 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



End Node

The end node is the final node of a workflow and is used to return the execution result of the workflow. Two types of parameters can be configured for the end node: input parameters and output parameters. They correspond to two return modes: text return and variable return.



Configure the node parameters as follows:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Input params: Input parameters support the reference and input types. An input parameter can be returned only after it is inserted as {{variable_name}} in the text box of the specified response.

Reply: You can edit the reply content in the text box. You can insert input parameters in the text box in the format of {{variable_name}} or directly return the input parameters. The final execution result of the workflow is returned based on the specified reply. Output parameters cannot be inserted in the specified reply.

Output params: Output parameters are returned as variables and support the reference and input types. After a workflow is executed, all output parameters are returned in JSON format. Output parameters apply to sub-workflows. If a workflow is bound to an agent and the workflow is triggered in a dialog, the model automatically summarizes the content in JSON format and replies to users in natural language. Output parameters cannot be referenced in a specified reply.

7.4.2.4.2 LLM Node

The LLM node allows you to call the capability of LLMs. You can configure a deployed model for the node, write prompts and set parameters to make models process specific tasks.

The LLM node is an **optional** node. If it does not need to be configured, skip this section.

To configure an LLM node, perform the following steps:

1. Drag the LLM node on the left to the canvas and click the node to open the node configuration page.
2. Configure the LLM node by referring to **Table 7-9**.

Table 7-9 LLM node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	Input params	Input parameters required for model processing. The values will be dynamically added to the prompt. The default input parameter name is **query**. • **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output parameters of the previous nodes contained in the workflow. – **literal**: transfers user-defined content to the LLM. In this mode, no matter what output content is generated by the previous node, the content will not be transferred to the LLM.

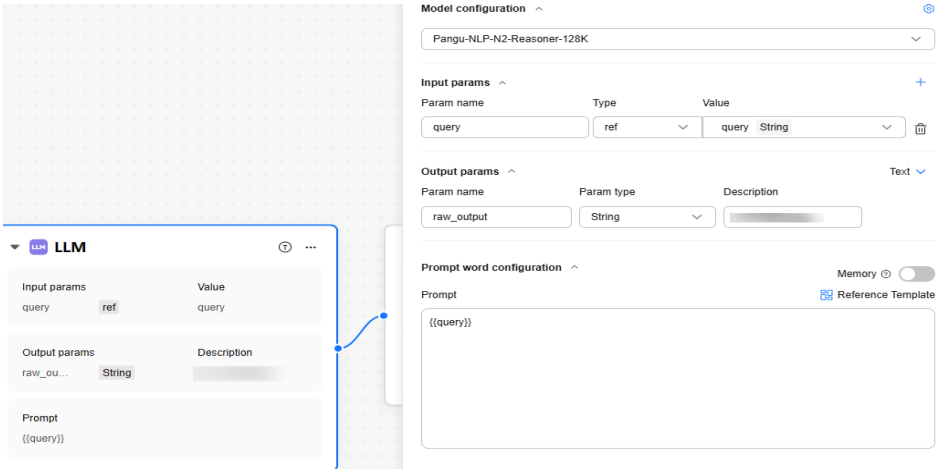
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Output params	This parameter is used to parse the output of the LLM node and provide the output for the subsequent node. • **Param name**: The length of the parameter name must be greater than or equal to one character. The character must be one of the following types: – Letters (A–Z and a–z) – Digits (0–9) – Special characters: underscores (_) **NOTE** The name of a user-defined output parameter cannot be the same as that of the built-in output parameter **rawOutput**. The LLM node has a built-in output parameter **rawOutput**, which indicates the original output of the node that is not parsed. The output can be directly referenced by the subsequent nodes connected to the LLM node. • **Param type**: type of an output parameter. The options are String, Integer, Number, and Boolean. • **Description**: description of the output parameter. • Output format: The supported output formats include Text, Markdown, and JSON. **Text**: original content output of the LLM. Only one parameter is supported. The default value is **raw_output**. The parameter name can be changed. **Markdown**: Select this option when the model is expected to output content in markdown format. Only one parameter is supported. The default value is **raw_output**. The parameter name can be changed. **JSON**: The model must respond in JSON format. Multiple parameters can be added.
Model configuration	Model selection	Select the LLM that has been configured in the model access module.
	Top P	During output, the model selects words with the highest probability until the total probability of these words reaches the Top P value. The Top P value can restrict the model to select these high-probability words, thereby controlling the diversity of output content. You are advised not to adjust this parameter together with the **Temperature** parameter.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Temperature	Controls the randomness of the generation result. A higher temperature makes the model output more diverse and innovative. A lower temperature makes the output more compliant with the instruction requirements, but reduces the diversity of the model output.
Prompt configuration	Prompt	Prompts instruct the model to complete tasks better. When configuring prompts, you can use the {{variable}} format to reference the parameters defined in the input parameters of the current node. The replaced content is transferred to the model.
	Memory	Whether to enable the memory function. After it is enabled, the content of multiple rounds of dialogs can be recorded. By default, it is disabled.

Figure 7-42 LLM node configuration example



3. After completing the configuration, click **OK**.
4. Connect the LLM node to other nodes.

7.4.2.4.3 Knowledge Repo Node

The Knowledge Repo node retrieves matched information from the specified knowledge base based on user input and returns the matching result in a list. You can select a knowledge base created by yourself. For details, see **Creating a Knowledge Base and Uploading Documents**.

To configure the Knowledge Repo node, perform the following steps:

1. Drag the Knowledge Repo node on the left to the canvas and click the node to open the node configuration page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. Connect the Knowledge Repo node to other nodes.
3. Configure the LLM node by referring to **Table 7-10**.

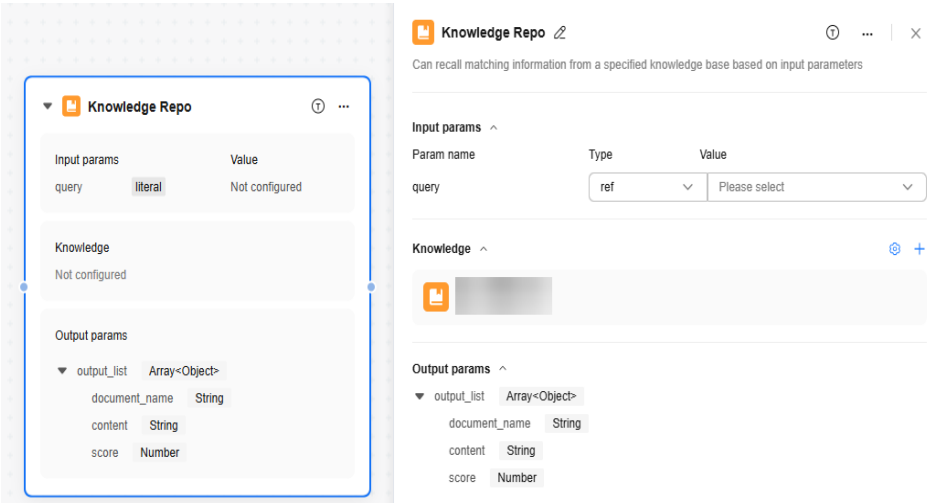
Table 7-10 Knowledge Repo node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	Input params	• **Param name**: There is only one input parameter. The parameter name is **query** and cannot be changed. The parameter value is a string, which represents the question to be searched for. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable value of a previous node that has been included in the workflow. The value must be of the String type. This option applies to the scenario where knowledge search questions need to be obtained from the output of a previous node. – **literal**: allows users to customize questions, which is applicable to scenarios where knowledge search questions are fixed.
	Output params	The output of the Knowledge Repo node is an object array. The parameter name is **output_list**, indicating all knowledge slices that meet the search requirements. The objects in the array have three attributes: • document_name: name of the knowledge document where the knowledge slice is located. • content: content of a knowledge slice. • score: matching score of a knowledge slice. Elements in output_list are sorted in descending order of scores. Subsequent nodes can reference the output parameter **output_list**. In this case, the full search results are obtained, including the document name, slice content, and score. You can also directly reference slice attributes, for example, **content**. In this case, the slice content of the first record in **output_list** is obtained.
Knowledge base	Knowledge	You can select a knowledge base created by users.
	Relevance Threshold	Search results whose scores are lower than the relevance threshold are filtered out. You can adjust the threshold based on the relevance score of the knowledge base hit test. The value ranges from 0 to 1.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Top K Recalls	Maximum number of segments recalled from the knowledge base. If the value is 5, segments whose scores are not in the top 5 will be filtered out. The value ranges from 1 to 50.
	FAQ Straightout threshold	FAQs in the knowledge base with scores higher than the threshold will be recalled. In this case, documents in the knowledge base are not searched. If no FAQ meets the threshold requirements, the system continues to search for documents in the knowledge base. You can adjust the threshold by referring to the FAQ hit test result in the knowledge base. The value ranges from 0 to 1.

Figure 7-43 Knowledge Repo node configuration example



4. After completing the configuration, click **OK**.

7.4.2.4.4 IntentDetection Node

The IntentDetection node is used to enable an application to understand the intent or purpose expressed by a user in natural language. It can be used in scenarios where user questions need to be classified or where different branches are processed for comprehensive functions.

The IntentDetection node performs inference and analysis on user input, matches predefined intent keyword categories, and directs the corresponding processing flow based on the matching result. This node is usually located in the front of the workflow.

The IntentDetection node can run in normal or advanced mode.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Normal mode: applies to scenarios where a small number of intents can be classified.
- Advanced mode: applies to scenarios where a large number of intents can be classified.

The IntentDetection node is an **optional** node. If it does not need to be configured, skip this section.

To configure the IntentDetection node, perform the following steps:

1. Drag the IntentDetection node on the left to the canvas and click the node to open the node configuration page.
2. Configure the IntentDetection node by referring to **Table 7-11**.

Table 7-11 IntentDetection node configuration description

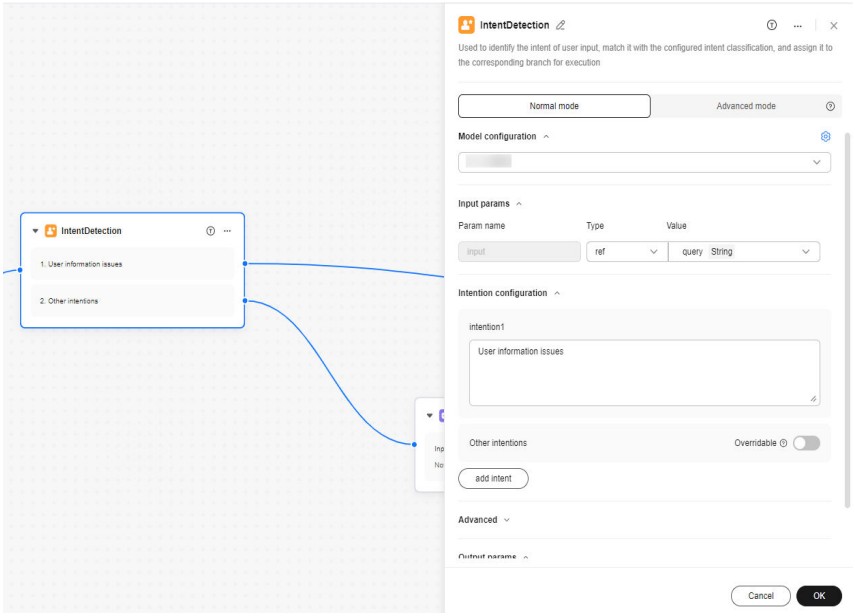
Configuration Type	Parameter Name	Description
Model configuration	Model selection	Model used for intent recognition. You can select any model accessed to the platform.
	Temperature	Controls the randomness of the generation result. A higher temperature makes the model output more diverse and innovative. A lower temperature makes the output more compliant with the instruction requirements, but reduces the diversity of the model output.
	Top P	During output, the model selects words with the highest probability until the total probability of these words reaches the Top P value. The Top P value can restrict the model to select these high-probability words, thereby controlling the diversity of output content. You are advised not to adjust this parameter together with the **Temperature** parameter.
Parameter configuration	Input params	• Param name: The default value is **input**, which is a fixed value and cannot be changed. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable values of the previous nodes contained in the workflow. – **literal**: The value can be customized.
	Output params	The **classification_id** and **name** parameters are fixedly output. • **classification_id**: sequence number of the recognized intent. • **name**: content of the recognized intent.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
Intention configuration	Intents	This area is used to configure intent keywords. You can add intents. The default intent categories are intention 1, intention 2, and so on. A maximum of 20 intents can be added. Enter the intent description in the text box. The description is a statement or keyword for the category, and is also used as a basis for inference and classification of the LLM.
	Other intents	Applies to scenario where the input intent of a user cannot be identified. If the predefined intent categories cannot be matched after inference and analysis, the subsequent process is executed by default in the branch corresponding to other intents.
Advanced settings	Prompt	Allows developers to modify prompts. If this configuration item is not set, the default value is used. The quality of prompts may affect the accuracy of the IntentDetection node.
	Historical dialogue rounds	Whether to enable the historical dialogue reference function. The default value is **0**, indicating that the dialogue history is not referenced. If this parameter is set to *N*, the content of *N*-round dialogues can be recorded.
	auxiliary identification	After **auxiliary identification** is enabled, intent recognition is preferentially performed based on the exact match of the category sample knowledge base to improve the classification capability of the IntentDetection node. • **intent sample knowledge**: To enable auxiliary identification, you need to create and configure a category sample knowledge base and configure the knowledge base. • **match threshold**: When the matching degree of a category sample is lower than the threshold, the default LLM is used for intent recognition and classification. The threshold ranges from 0 to 1.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-44 IntentDetection node configuration example



3. After completing the configuration, click **OK**.
4. Connect the IntentDetection node to other nodes.

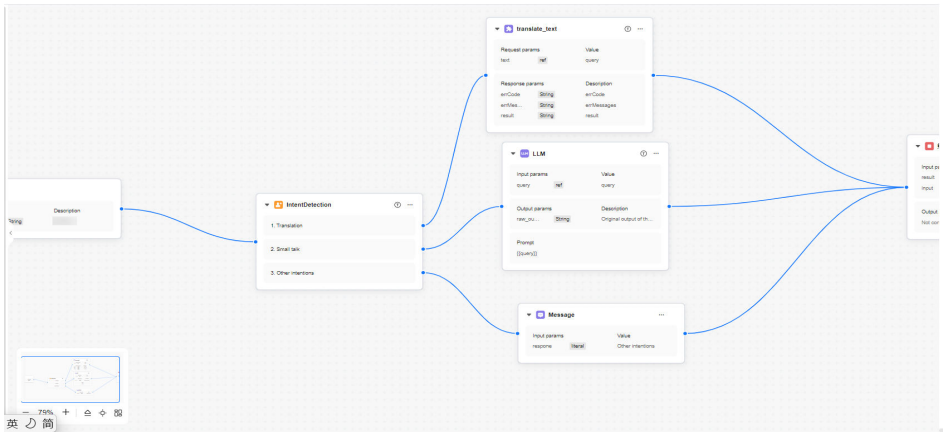
Examples:

For example, in a workflow that provides comprehensive functions and processes user questions in different branches, the IntentDetection node classifies user input and transfers the input to different functional modules for processing.

For example, for a workflow that provides the translation function, the node configuration is as follows:

- IntentDetection node: The workflow classifies user queries into two categories: translation and chat. The category description of intent 1 is added to the intent configuration of the node as text translation. The category is followed by the Translation Plug-in node to implement the translation function. The category description of intent 2 is chat, the category is transferred to the LLM node to implement the chat function. Other default intent categories are transferred to the Message node. The default reply content is configured in the Message node to implement the back-stop reply in the scenario where the intent is not identified.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



Advanced mode

If there are more than 100 intent branches, you are advised to use the advanced mode. In this mode, you can configure intent branch information on an independent page and complete service configuration by selecting the interaction mode of sub-workflows.

Procedure:

1. Configure an intent package.
   a. Choose **Configuration Management** > **Intent Management** from the navigation pane and click **Create Intent Package**.
   b. Click ＋ to add an intent category to the intent package. The category information includes the name and example.

   Figure 7-45 Configuring an intent package



2. Drag the IntentDetection node on the left to the canvas and click the node to open the node configuration page. Switch to the advanced mode.
3. Complete the configuration by referring to the IntentDetection mode configuration description.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

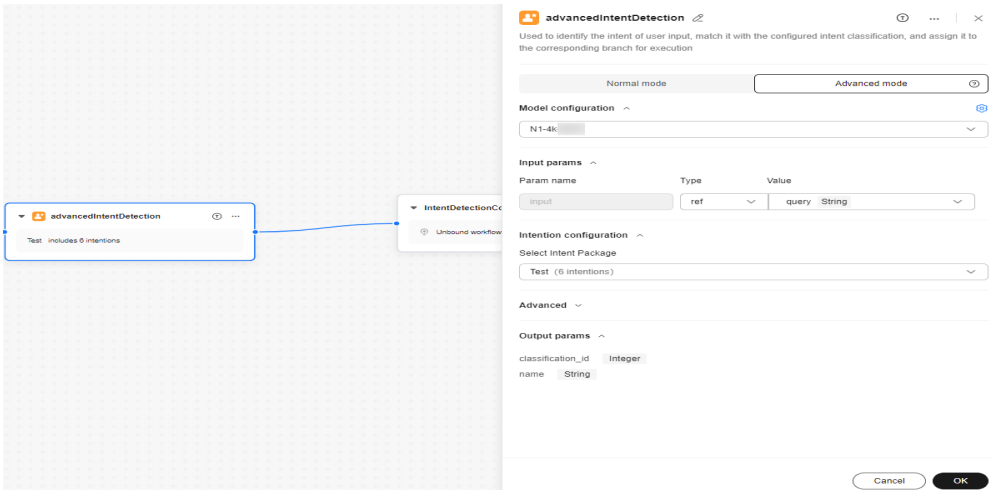
Table 7-12 IntentDetection node configuration description

Configuration Type	Parameter Name	Description
Model configuration	Model selection	Model used for intent recognition. You can select any model deployed on the platform.
	Temperature	Controls the randomness of the generation result. A higher temperature makes the model output more diverse and innovative. A lower temperature makes the output more compliant with the instruction requirements, but reduces the diversity of the model output.
	Top P	During output, the model selects words with the highest probability until the total probability of these words reaches the Top P value. The Top P value can restrict the model to select these high-probability words, thereby controlling the diversity of output content. You are advised not to adjust this parameter together with the **Temperature** parameter.
Parameter configuration	Input params	• Param name: The default value is **input**, which is a fixed value and cannot be changed. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable values of the previous nodes contained in the workflow. – **literal**: The value can be customized.
	Output params	The **classification_id** and **name** parameters are fixedly output. • **classification_id**: sequence number of the recognized intent. • **name**: content of the recognized intent.
Intention configuration	Intent package	Select the configured intent package.
Advanced settings	Prompt	Allows developers to modify prompts. If this configuration item is not set, the default value is used. The quality of prompts may affect the accuracy of the IntentDetection node.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

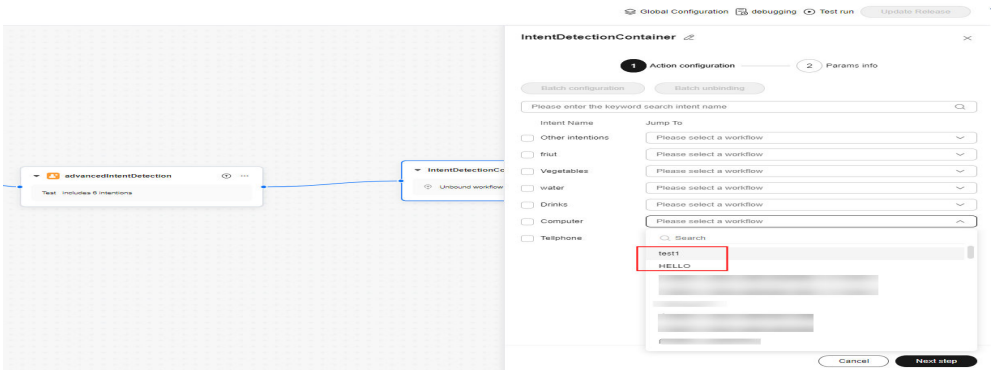
Configuration Type	Parameter Name	Description
	Memory	Whether to enable the memory function. By default, this function is disabled. If it is enabled, the content of multiple rounds of dialogs can be recorded.

Figure 7-46 IntentDetection node configuration example



4. After completing the configuration, click **OK**.
5. Click the intent action node and configure the processing logic corresponding to the branch.

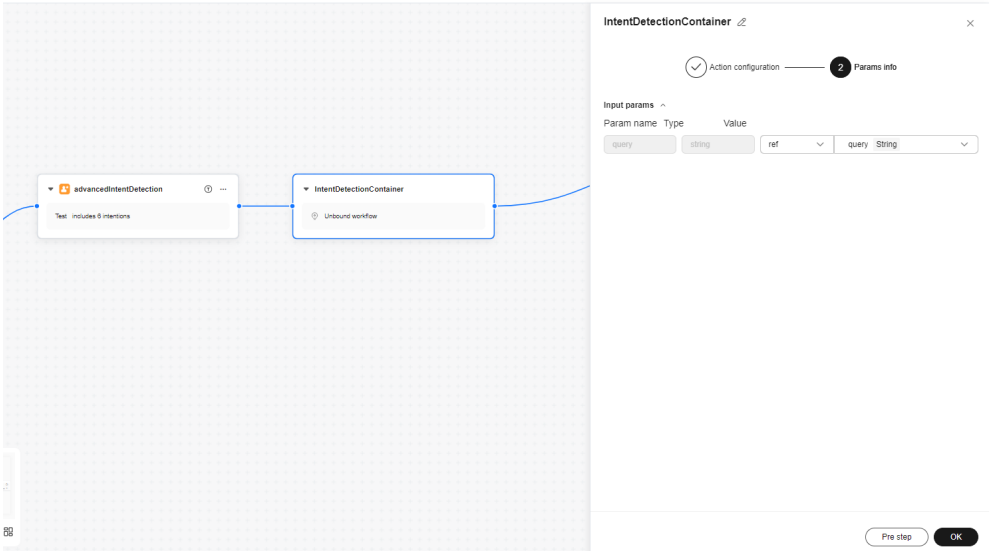
Figure 7-47 Configuring the processing logic



6. Set the input parameters of the sub-workflow.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-48 Setting input parameters



7. Click **OK**.
8. Connect the intent action node to other nodes.

7.4.2.4.5 Plugin Node

The Plugin node is the core component for calling third-party capabilities in a workflow. As an important carrier for function extension, this node can invoke this node to execute specific function tasks. Each plug-in is essentially a set of API tools that are encapsulated in a standardized manner. Plug-ins provide plug-and-play modular services, expand the capability boundary of workflows, and complete more complex tasks.

There are preset plug-ins and personal plug-ins.

- Preset plug-ins: The code interpreter plug-in is preset on the platform to execute the Input code and obtain the execution result. Developers can directly add plug-ins to workflows or applications to enrich their capabilities.
- Personal plug-ins: The platform allows you to create custom plug-ins. APIs can be quickly created as plug-ins through configuration for workflows to use.

The Plugin node is an **optional** node. If it does not need to be configured, skip this section.

Plugin Node Configuration Example

1. Drag the Plugin node on the left to the canvas. On the **Personal plugin** or **Pre installed plugins** tab page, click + to add a plug-in to the canvas.

   Pre-installed plug-ins are built-in plug-ins of the platform, and personal plug-ins are user-defined plug-ins. For details about how to create a plug-in, see **Creating a Plug-in**.
2. Connect the Plugin node to other nodes.
3. Click the added Plugin node in the canvas and configure the node by referring to **Table 7-13**.

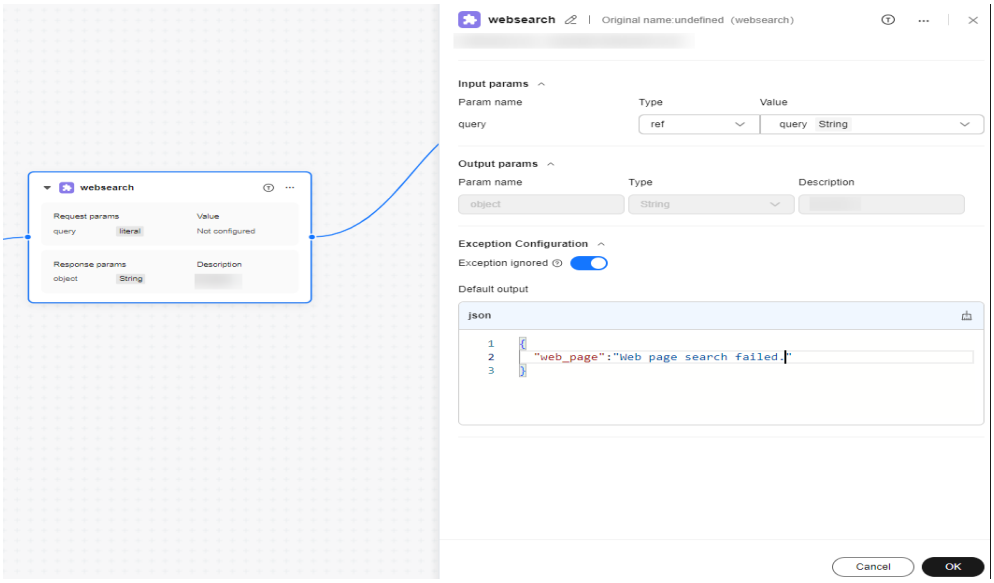
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-13 Plugin node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	Input params	• **Param name**: imported from the plug-in metadata. You do not need to manually add a parameter. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable value of a previous node that has been included in the workflow. This option applies to the scenario where plug-in input parameters need to be obtained from the output of a previous node. – **literal**: You can customize the value. This type applies to the scenario where the plug-in input parameter value is fixed.
	Output params	All output parameter information is automatically imported from the plugin metadata. You do not need to manually modify the information.
Exception Configuration	-	The configuration of the **Exception ignored** switch is used as the default configuration of the Plugin node. When the plug-in execution is abnormal, the default configuration can be returned to prevent the entire workflow from failing to run. • When this function is disabled, exceptions are not ignored. • When this function is enabled, set **Default output**. The parameters in **Default output** must be the same as those in **Output params** of the plug-in. When the Plugin node is running properly, the configuration does not affect subsequent nodes. If an exception occurs when a workflow runs to the Plugin node, the workflow does not stop and continues to run subsequent nodes. If a subsequent node references the output content of the Plugin node, the default output content is used.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-49 Plugin node configuration example



4. After completing the configuration, click **OK**.

Example: Creating a Plug-in Using a Deployed Model

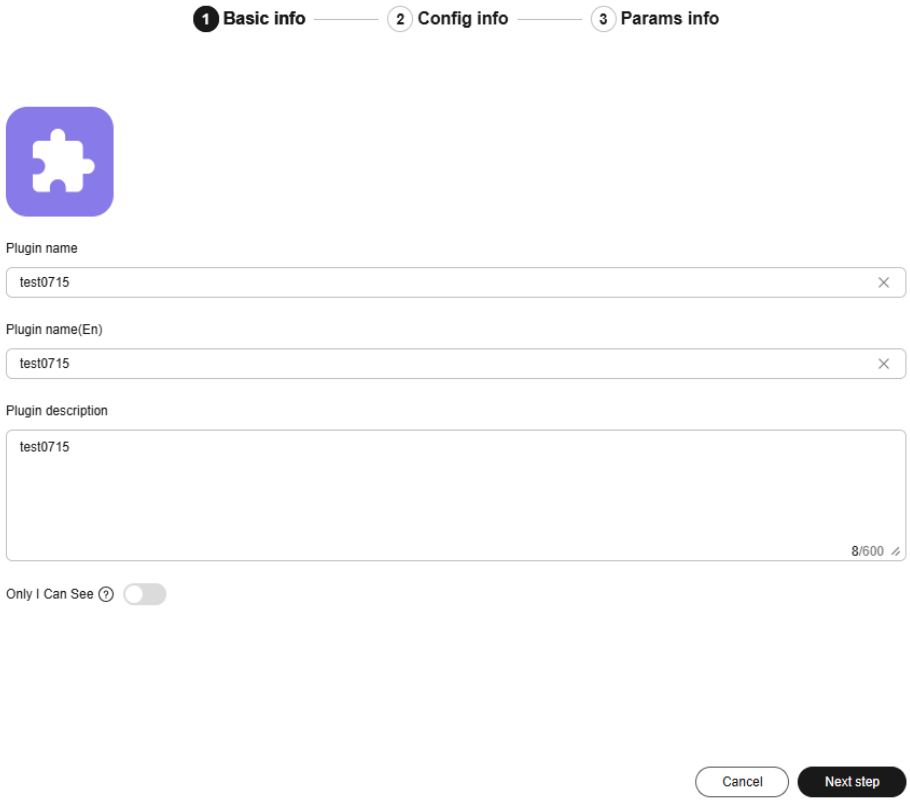
Step 1 Deploy a service, for example, **a DeepSeek model**.

Obtain the API request URL. For details, see **Obtaining the API Request URL of the Third-Party Model**.

Step 2 Create a plugin on the platform and enter basic information. For details, see **Creating a Plug-in**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

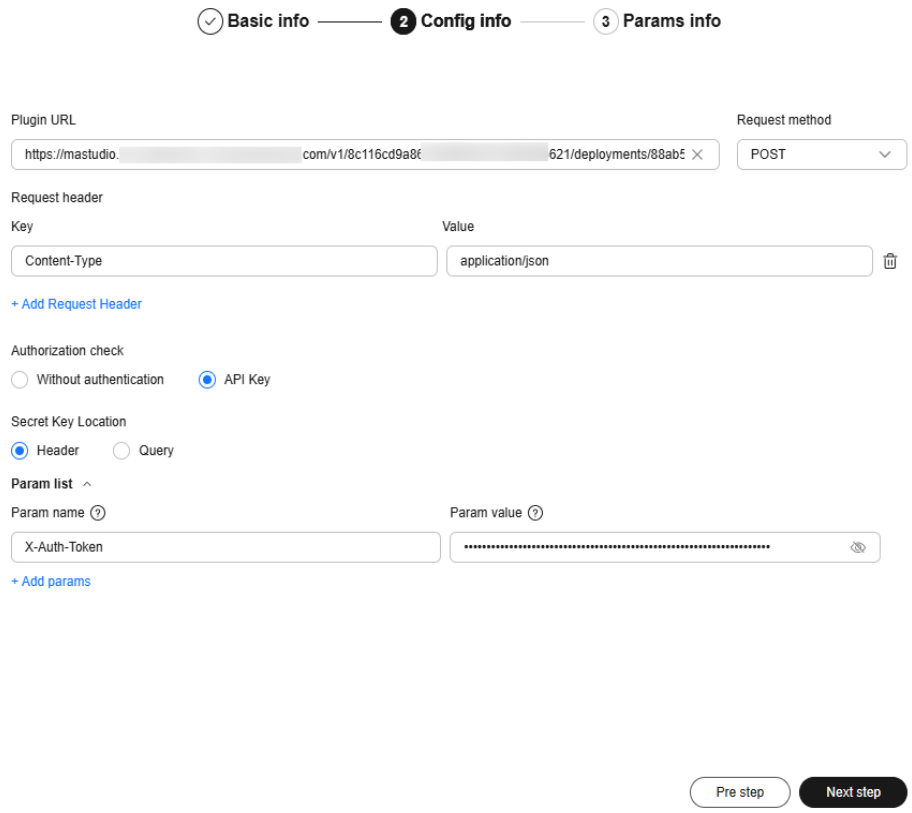
Figure 7-50 Basic information



Step 3 Configure the **Plugin URL** and other request information.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

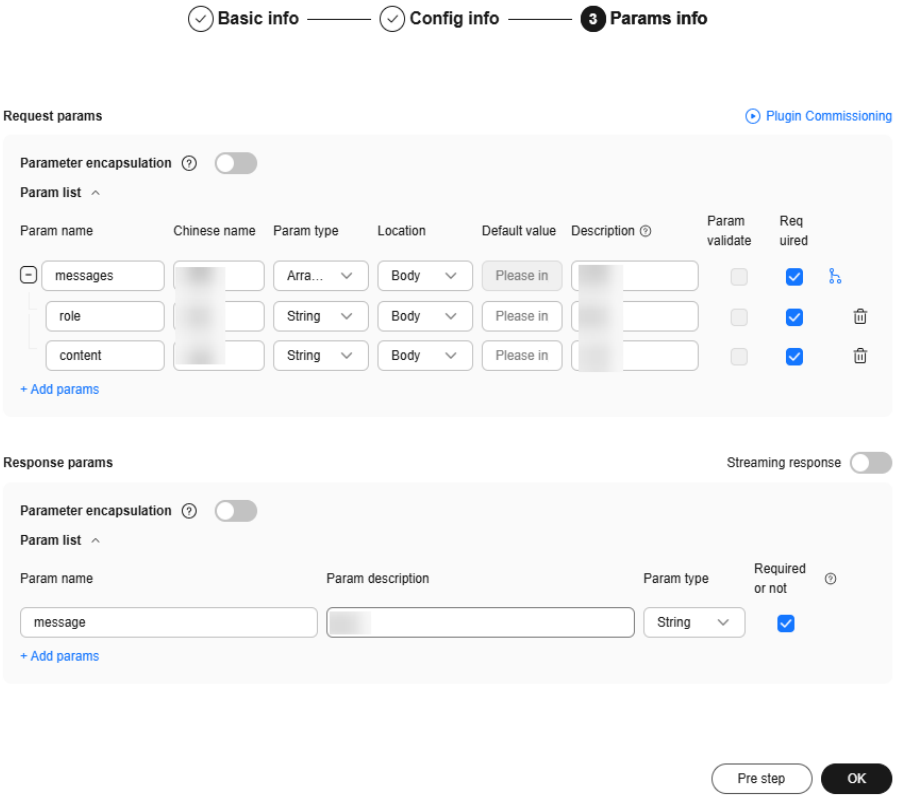
Figure 7-51 Configuration



Step 4 Set parameters based on the API information and click **Plugin Commissioning**. After confirming the information, click **OK** to create the plugin.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-52 Parameters

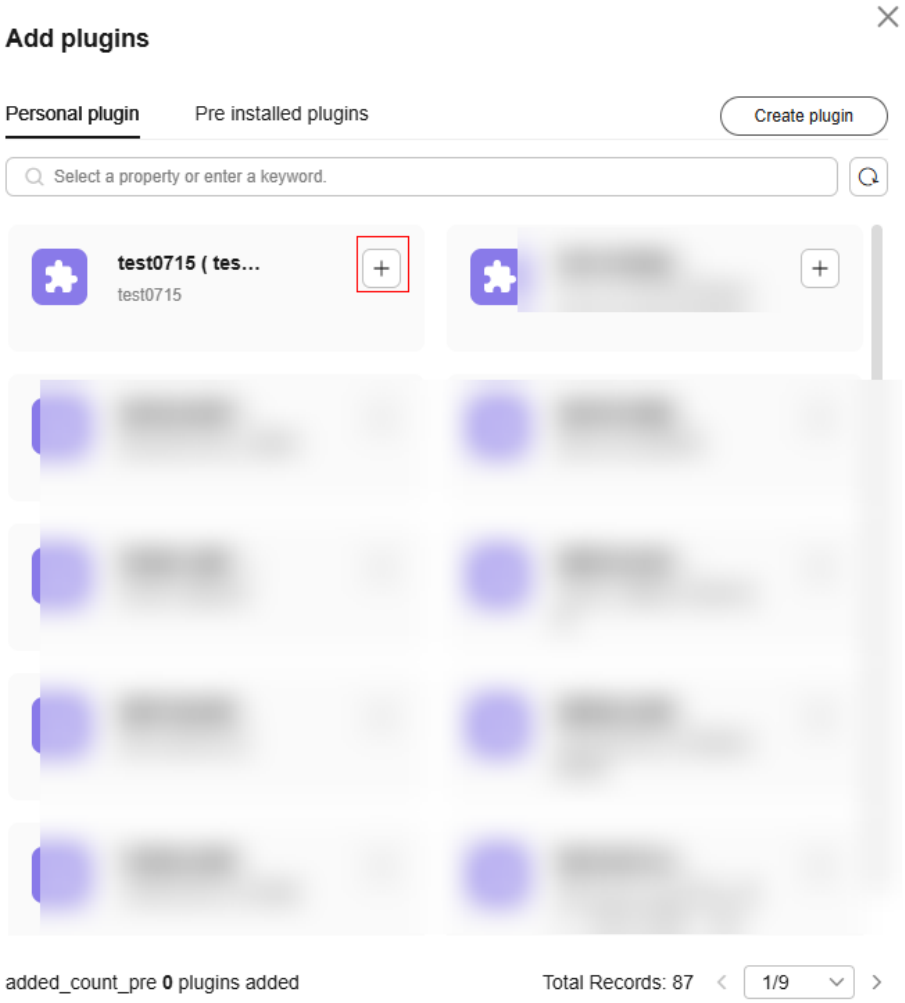


Step 5 Use this plug-in in the workflow.

Drag the Plugin node on the left to the canvas. On the **Personal plugin** tab page, click + to add a plug-in to the canvas.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-53 Using the plug-in



----End

7.4.2.4.6 Branch Node

A Branch node is an IF-ELSE node that provides the multi-branch condition judgment capability. It is used to design branch processes and implement the logic judgment function.

Each condition branch of the Branch node supports multiple judgment conditions (AND and OR) and multiple condition branches.

When parameters are input to the Branch node, the node checks whether the input meets the preset conditions in each condition branch one by one. If the input meets the preset conditions, the workflow process after the corresponding branch is executed. If no branch meets the preset conditions, the workflow branch corresponding to ELSE is executed.

The Branch node is **optional**. If it does not need to be configured, skip this section.

To configure the Branch node, perform the following steps:

1. Drag the Branch node on the left to the canvas and click the node to open the node configuration page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

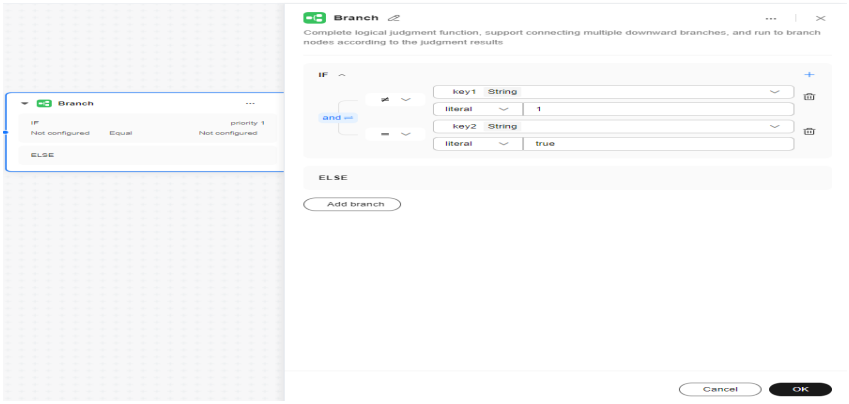
2. Configure the Branch node by referring to **Table 7-14**.

Table 7-14 Branch node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	IF	The IF branch consists of **[Judgment parameter comparison Condition comparison Parameter]**, which forms a condition expression. • Judgment parameter: left part of the condition expression. You need to select the output parameter from the previous node. • Comparison condition: middle part of a condition expression. Currently, the following comparison conditions are supported: length greater than, length greater than or equal to, length less than, length less than or equal to, equal to, not equal to, including, not including, empty, and not empty. Comparison conditions displayed vary depending on judgment parameter types. • Comparison parameter: right part of a condition expression. It supports the reference and input types. – **ref**: You can select the output variable values of the previous nodes contained in the workflow. – **literal**: The value can be customized. • **Add conditions**: Click + to add multiple condition expressions to the current condition branch. The multiple condition expressions are connected by AND or OR. – Click **and** or **or** to switch the operation logic of the branch expression.
	ELSE	Determines whether to execute subsequent workflows in the default branch if no preset condition branch is met. If no branch meets the conditions, the default branch is used.
	Add branch	You can add a new ELSE IF branch. The configuration method of the new branch is the same as that of the IF branch.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-54 Branch node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Branch node to other nodes.

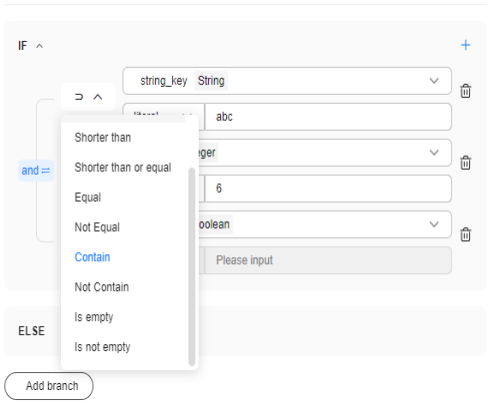
Examples:

The following uses the String, Integer, and Boolean types as an example to describe how to define parameters of the three types in the Start node to simulate the input parameters of the Branch node and implement logical judgment on parameters of different types under different conditions.

The node configuration is as follows:

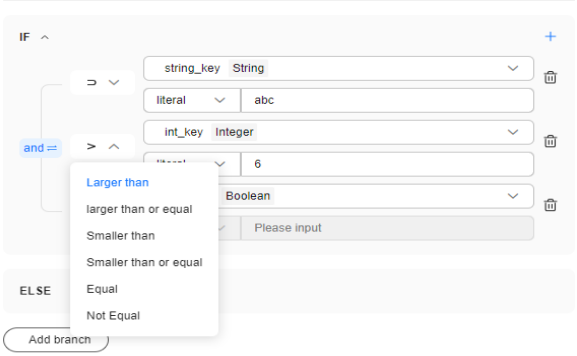
- Start node: Define three types of parameters: string_key of the string type, int_key of the integer type, and bool_key of the Boolean type.
- Branch node: Add three conditions to the IF branch. The parameters of the condition expression reference the preceding three types of parameters of the Start node. The comparison conditions displayed vary according to the parameter type.

  For example, for the String type, the comparison conditions are about the length, inclusion, and emptiness of the string. In the example, whether string_key includes "abc" is checked.



  The Integer type is used to determine whether the value meets the "greater than", "less than", or "equal to" conditions. In the example, whether int_key is greater than 6 is checked.

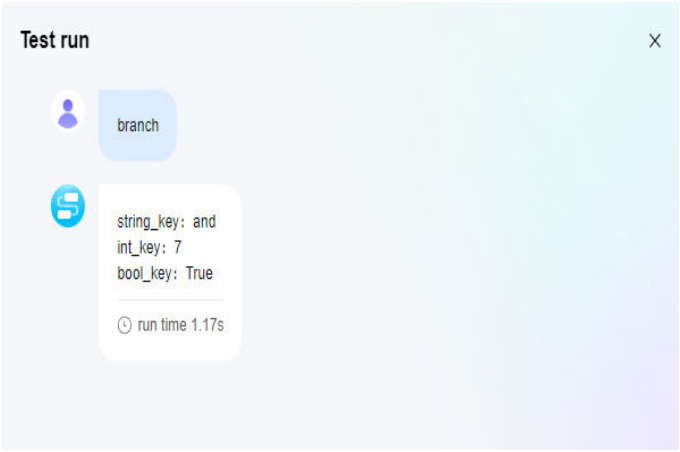
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



The Boolean type is used to determine whether the value is true or false based on the preset condition. In the example, whether bool_key is true is checked.



Click **Test run** and enter the following configuration: string_key: abcd, int_key: 7, and bool_key: true. View the execution result.



7.4.2.4.7 Code Node

The Code node allows you to write Python code to process complex logic such as text and generate expected return values.

The Code node is an **optional** node. If it does not need to be configured, skip this section.

To configure the Code node, perform the following steps:

1. Drag the Code node on the left to the canvas and click the node to open the node configuration page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. Configure the Code node by referring to **Table 7-15**.

Table 7-15 Code node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	Input params	Configure the input parameters required for code running. • **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output parameters of the previous nodes contained in the workflow. – **literal**: The value can be customized.
	Output params	Configure parameters that need to be output after the code is executed. The parameters must be the same as those returned by **return**. • **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Param type**: output parameter type. The options are String, Integer, Number, Boolean, and more. • **Description**: description of the output parameter. • **Required**: whether the current output parameter is mandatory.

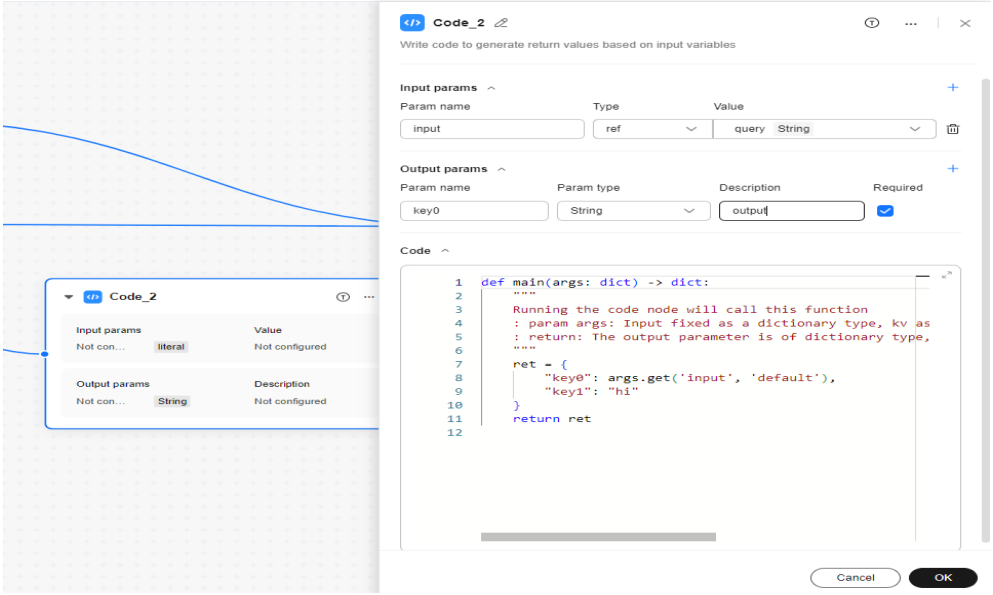
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description

Code	-	You can compile Python code snippets. Only the Python system library is supported, and dependency packages cannot be imported. The following is an example of code configuration:
		• Sample code for text stitching
		`def main(args: dict) -> dict:`
		`# Define a variable named input1 in the input parameters.`
		`input1 = args.get('input1')`
		`# Define a variable named input2 in the input parameters.`
		`input2 = args.get('input2')`
		`res = {`
		`# Define a variable named res in the output parameters.`
		`"res": input1 + input2,`
		`}`
		`return res`
		• Sample code for mathematical calculation
		`def main(args: dict) -> dict:`
		`# Define a variable named input1 in the input parameters.`
		`input1 = args.get('input1')`
		`try:`
		`input1 = int(input1)`
		`return {`
		`# Define res in the output parameters.`
		`'res': input1 * input1`
		`}`
		`except Exception as e:`
		`return {`
		`# Define res in the output parameters.`
		`'res': "The input type is incorrect or the number exceeds the upper limit."`
		`}`
		NOTE
		Do not change the function definition in the first line when writing code.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-55 Code node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Code node to other nodes.

7.4.2.4.8 Message Node

The Message node provides the message output capability in the intermediate process. You can define a piece of text content. The content will be sent to users during workflow execution.

Generally, after a workflow is executed, the final execution result is output through the End node. If the result of an intermediate node during workflow execution needs to be output, you can use this node.

The Message node is an **optional** node. If it does not need to be configured, skip this section.

To configure the Message node, perform the following steps:

1. Drag the Message node on the left to the canvas and click the node to open the node configuration page.
2. Configure the Message node by referring to **Table 7-16**.

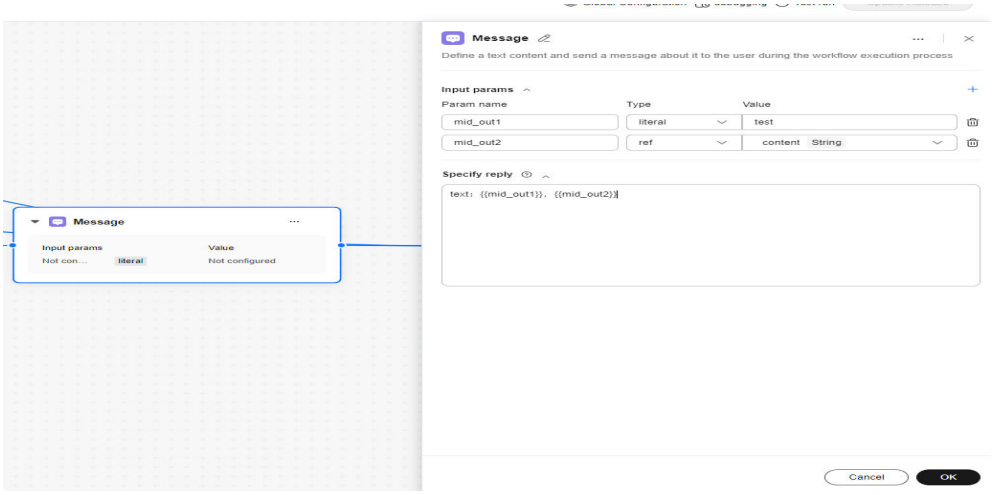
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-16 Message node configuration

Configuration Type	Parameter Name	Description	Configuration Example (Scenario-based)
Parameter configuration	Input params	• **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable values of the previous nodes contained in the workflow. – **literal**: The value can be customized.	When the workflow runs to this node, if you want to send a specified message that contains the input or output parameters of the previous node to the user, you can configure the parameters in **ref** mode. If you want to send a specified message with fixed content to the user, you can configure the parameters in **literal** mode.
Specify reply	-	You can write specified reply information and insert variables in the {{Parameter name}} format. The reply is sent to users when the workflow reaches the node.	For example, if you want to send the output parameter **content** of the Knowledge Repo node to users during workflow execution, you can add the input parameter **rep_out**, reference **content**, and write {{rep_out}} in the specified reply.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-56 Message node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Message node to other nodes.

7.4.2.4.9 Questioner Node

The Questioner node provides developers with the function of collecting information required by user questions. This node is executed cyclically until all required information is collected.

The Questioner node is an **optional** node. If it does not need to be configured, skip this section.

To configure the Questioner node, perform the following steps:

1. Drag the Questioner node on the left to the canvas and click the node to open the node configuration page.
2. Configure the Questioner node by referring to **Table 7-17**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-17 Questioner node configuration

Configuration Type	Parameter Name	Description
Mode preference	-	**Effect priority**: In this mode, the time augmentation and reflection functions are enabled, which improves the parameter improvement success rate but increases the latency. • Time augmentation: The time and date in the natural language can be extracted as the time and date in the YYYY-MM-DD HH:MM:SS format. For example, if the time is 12:30 tomorrow, the time is extracted as 2024-04-13 13:30:00. • Reflection: After data is extracted, the model determines whether the data is correctly extracted and whether the format meets the requirements. If the format does not meet the requirements, the model makes some corrections. For example, if the expected extracted data is a phone number, the user enters "I don't remember the phone number", and the extracted data is 189******, then the model determines that the extracted data is incorrect and continues to ask the user more questions. Speed priority: The latency is the lowest in this mode. The parameter improvement success rate may not be ensured. In this mode, the time augmentation and reflection functions are disabled.
Model configuration	Model selection	Select the model to be executed on this node. You can set parameters such as the generation diversity of the model on this node to make the model effectiveness meet your expectation. The Questioner model is used to receive the natural language of the user and extract the output parameters configured by the user. When **Effect priority** is selected, the Questioner node is also used to extract the result for reflection and correction.
	Top P	During output, the model selects words with the highest probability until the total probability of these words reaches the Top P value. The Top P value can restrict the model to select these high-probability words, thereby controlling the diversity of output content. You are advised not to adjust this parameter together with the **Temperature** parameter.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Temperature	Controls the randomness of the generation result. A higher temperature makes the model output more diverse and innovative. A lower temperature makes the output more compliant with the instruction requirements, but reduces the diversity of the model output.
Parameter configuration	Input params	Set the parameters to be added to the question. The parameter values can reference the output parameters of the previous node or be set to fixed text content. Multiple parameters can be referenced. • **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. Example: The input parameter is **pre_assigned_meeting_rooms**. You want the user to select one from the specified options. The follow-up question is set to "The following meeting rooms are available: {{pre_assigned_meeting_rooms}}. Please select the meeting room you want to reserve." • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable of the previous node of the node in the workflow as the value. – **literal**: You can enter the variable value.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Output params	This parameter is used to parse the output of the LLM node and provide the input for the subsequent node. Multiple parameters can be extracted. • **Default output** – USER_RESPONSE: original user output – STATUS: extraction status 0: The extraction is successful, and the user has not confirmed the result. 10: The extraction is successful, and the user has confirmed the result. 100: Some parameters are obtained. The user proactively interrupts the operation. An error is reported for the parameters that have been submitted, and the parameters that have not been submitted are left empty based on the format. 101: Some parameters are obtained. The number of cycles exceeds the preconfigured number of rounds. The parameters that have not been submitted are left empty based on the format. 201: The LLM failed to be called. 202: An error occurred in the reflection module. • **parameter extraction**: After this function is enabled, you can add parameters to be extracted. The configurable attributes of the parameters are as follows: – **Param name**: The value can contain only letters, digits, underscores (_), and hyphens (-). – **Chinese name**: It cannot be empty. – **Type**: type of an output parameter. The options are String, Integer, Number, and Boolean. – **Default value**: default value of the output parameter. If the LLM cannot extract the parameter and the maximum number of reply rounds is reached, the default value is used. – **Description**: description of the output parameter. – **check**: After this function is enabled, you can customize parameter verification rules to verify the standardization of output parameters. The rule includes the parameter name, verification type, and verification rule.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

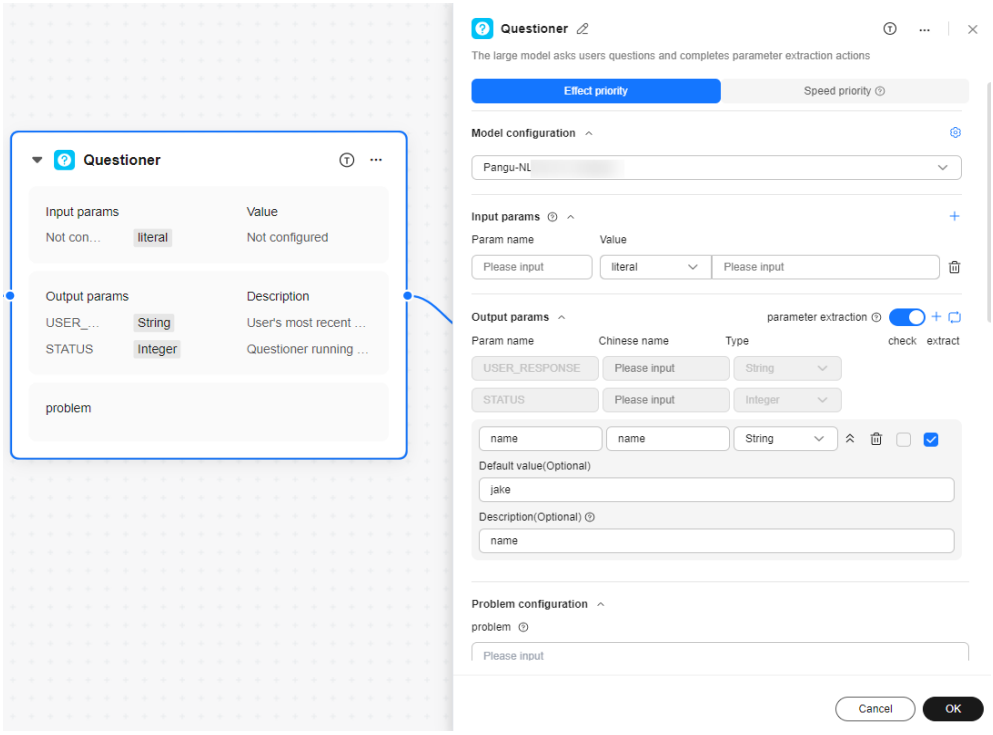
Configuration Type	Parameter Name	Description
		– **extract**: If this function is enabled, the parameter must be extracted or the default value is used if it is configured. If this function is disabled, the parameter can be left empty. • **Reference plug-in**: Extracted parameters may be used by plug-ins. By referencing a plug-in, you can import the parameter information and verification information of the plug-in to improve the configuration efficiency.
Problem configuration	problem	If this parameter is not set, the LLM automatically generates a question containing all question keywords based on the output parameter description. Example: What is your name? Input parameters can be used in questions using the Jinjia syntax. Example: Which class are you in? Available classes are {{classes}}. (The **classes** variable is preconfigured in the input parameters first.)
	Maximum number of reply rounds	This parameter specifies the maximum number of interaction rounds with the model. If no parameter is extracted when the maximum number of interaction rounds is reached, the questioner exits.
Advanced settings	Allow user to exit interaction	After this function is enabled, if a user expresses the intention of ending the dialog with the questioner, the system automatically ends the current question and switches to the End node.
	user_confirmation_extraction_param	After this function is enabled, users can confirm the questioner parameter extraction result.
	Extract constraints	Provides additional constraints of the LLM for more accurate parameter extraction, for example, specifying the format requirements of the extracted parameters. Example: The user wants to extract the phone number tel_number. The constraint configuration shows that the value of tel_number must contain 11 digits.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Follow up mode	The follow-up mode is used to configure the generation mode of the parameter follow-up questions returned by the system during multiple interactions. • **Default**: The default built-in follow-up question template is used to generate follow-up questions. The content of each follow-up question is the same. • **Intelligent questioning**: The large model is used to generate follow-up questions with good semantics and rich expressions. The content of each follow-up question is rich and diversified. • **Custom questioning**: Follow-up questions are generated based on the user-defined template configuration. {unextracted_cn_field_names} cannot be modified or deleted. The content of each follow-up question is the same. Example: Extract the name and age parameters. • **Default**: Please provide your name and age. • **Intelligent questioning**: Hello, we need to obtain your name and age. (The content is generated by the model and is not fixed.) • **Custom questioning**: The customized question asking template is configured as follows: Please provide the following information: {unextracted_cn_field_names}. Please provide the following information: name and age.
	Example configuration	Provide an expected parameter extraction example of the LLM to enhance the LLM's understanding of the parameter extraction scenario. Template: Input query: I want to fly to Hohhot for training. Parameters extracted: {"location": "Hohhot", "traveltool": "Aircraft"}
-	External API rewriting	After this function is enabled, an external API is called to rewrite user input questions before the LLM Q&A module is called. The rewriting API must meet the following requirements: 1. The POST request is sent, and no authentication information is required. 2. The input body contains only the **text** field, for example, ("text": "Content to be rewritten"}. 3. The returned structure contains only the **text** field, for example, {"text": "Rewritten content"}.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-57 Questioner node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Questioner node to other nodes.

7.4.2.4.10 Loop Node

The Loop node provides the capability of executing nodes cyclically. You can configure the nodes to be executed cyclically in the loop body. You can orchestrate nodes in the loop body to execute processing tasks for multiple times.

A loop is a common control mechanism used to repeatedly execute a series of tasks until a certain condition is met. You can add a Loop node for a workflow if you need to repeatedly perform some operations or process a group of data in a loop.

The Loop node is an **optional** node. If it does not need to be configured, skip this section.

To configure the Loop node, perform the following steps:

1. Drag the Loop node on the left to the canvas, drag other nodes that need to be executed cyclically to the loop body canvas, and orchestrate the nodes. (The execution in the loop must start from the loop input node and the output must be connected to the loop output node. Interactive nodes are not supported currently.)
2. Configure the Loop node by referring to **Table 7-18**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-18 Loop node configuration

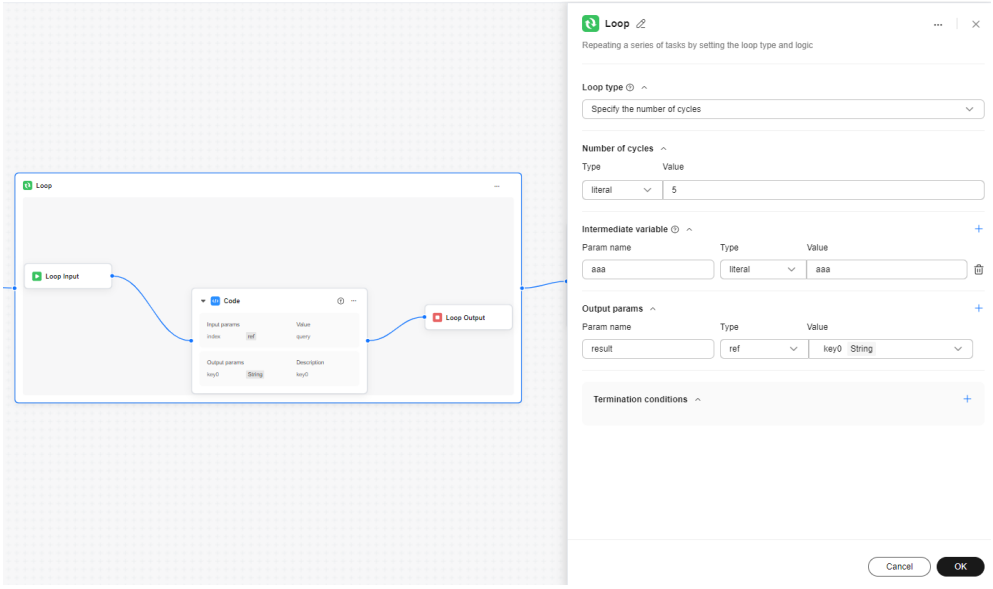
Configuration Type	Parameter Name	Description
Loop type	Using array loop	The array loop is similar to the for syntax structure in the programming language. The traversal loop is used to traverse a known sequence and perform a series of same steps on each element in the sequence. When using the array loop, you need to specify the value of **arr_loop_var**. This parameter can only reference the output of the upstream node and must be in array format. When a Loop node is executed in array loop mode, the number of loops depends on the length of the array referenced by the loop array. When an array loop is used, the Loop node traverses each element in the array. In each loop, the element in the current loop is assigned to the built-in variable. Built-in variables can be used only in Loop nodes. Currently, the following built-in variables are supported: • **item**: array element, that is, the array element that is looped to. • **index**: array index. index+1 indicates the round of the current loop.
	Specify the number of cycles	This loop type is usually used in scenarios where data is processed in batches or in sequence. You also need to set **Number of cycles**. The default number of cycles is 5. You can set a value ranging from 1 to 1000. Reference: For a turn-based game, set the number of cycles to 3 if the winner is the one who wins two out of three rounds. For a web crawler, set the number of cycles to 1000 if the crawler needs to crawl information about the first 1000 offerings.
Variable configuration	loop_array	This parameter can be configured only when the array loop is used. The name is fixed to **arr_loop_var**. Only the output of the upstream node can be referenced.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Intermediate variable	An intermediate variable can be set for a Loop node. This variable can be used for each loop. An intermediate variable is usually used together with the Variable Assignment node in the loop body. After each loop ends, a new value is set for the intermediate variable and the new value is used in the next loop. • **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable values of the previous nodes contained in the workflow. – **literal**: The value can be customized.
	Output params	The output parameters of the Loop node can be set to the execution result set of the loop body, meaning that after all elements in the array are executed, the execution results of all loops are packaged and output to the downstream. It can also be set to the value of an intermediate variable.
Termination conditions	Expression	The branch consists of **[Judgment parameter comparison Condition comparison Parameter]**, which forms a condition expression. • Judgment parameter: left part of the condition expression. You need to select the output parameter from the previous node. • Comparison condition: middle part of a condition expression. Currently, the following comparison conditions are supported: length greater than, length greater than or equal to, length less than, length less than or equal to, equal to, not equal to, including, not including, empty, and not empty. Comparison conditions displayed vary depending on judgment parameter types. • Comparison parameter: right part of a condition expression. It supports the reference and input types. – **ref**: You can select the output variable values of the previous nodes contained in the workflow and the memory variables in the global configuration. – **literal**: The value can be customized. • **Add conditions**: Click + to add multiple condition expressions to the current condition branch. The multiple condition expressions are connected by AND or OR. – Click **and** or **or** to switch the operation logic of the branch expression.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-58 Loop node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Loop node to other nodes.

7.4.2.4.11 Variable Assignment Node

Variable Assignment Node Configuration

The Variable Assignment node is used to assign variable values to writable variables, achieving dynamic updating and transmission of data. With this node, the workflow can make corresponding processing and decisions based on real-time data. This node can also be used within a Loop node. The Variable Assignment node assigns specific values to intermediate variables, achieving dynamic updating and transmission of data in the loop.

To configure the Variable Assignment node, perform the following steps:

1. Drag the Variable Assignment node on the left to the canvas and click the node to open the node configuration page.
2. Connect the Variable Assignment node and other nodes.
3. Configure the Variable Assignment node by referring to **Table 7-19**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-19 Variable Assignment node configuration

Configuration Type	Parameter Name	Description
Configuration of the Variable Assignment node outside the Loop node	Variable Assignment	**Variable name**: The variable name can be referenced only in the global configuration. The value can be the reference or input type. **Type** and **Value**: **Type** can be set to **ref** and **literal**. • **ref**: You can select the output variable values of the previous nodes contained in the workflow and the memory variables in the global configuration. • **literal**: The value can be customized.
Configuration of the Variable Assignment node within the Loop node	Variable Assignment	The Variable Assignment node can be referenced in the loop body. Only the value of the intermediate variable in the loop body can be changed. Only intermediate variables can be selected as the variables to be assigned. The value can be the reference or input type. This setting is applicable to the scenario where intermediate variables are dynamically updated during a loop and parameters are transferred in the customized loop logic. **Type** and **Value**: **Type** can be set to **ref** and **literal**. • **ref**: Select this option if the value of the intermediate variable needs to reference the output of the upstream node. You can select the output variable value of the preceding node that has been included in the workflow and the built-in variables of the loop body, including index, item (array loop), and intermediate variable. This option applies to the scenario where the value of the intermediate variable is changed to a variable during the loop. • **literal**: You can customize the value. This option applies to the scenario where the value of the intermediate variable is changed to a fixed value during the loop.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-59 Variable Assignment node configuration example

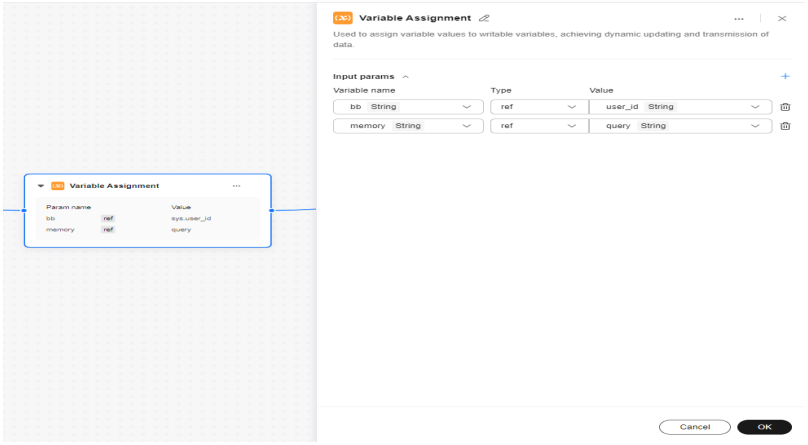
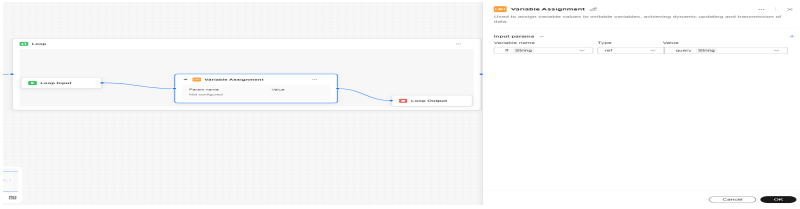


Figure 7-60 Example of configuring the Variable Assignment node within the Loop node



4. After completing the configuration, click **OK**.
5. Connect the Variable Assignment node and other nodes.

7.4.2.4.12 Aggregation Node

The Aggregation node can aggregate the output of multiple branches, facilitating unified configuration of subsequent nodes.

If multiple branches are designed in a workflow, a node is required to summarize the output results of all branches. In this scenario, you can use the Aggregation node to aggregate the output variables of multiple branches. The Aggregation node reads the first non-null value from multiple branches for the downstream nodes to use and operate. You do not need to process the output results of the branches that are not running, simplifying data flow management.

The Aggregation node is an **optional** node. If it does not need to be configured, skip this section.

To configure the Aggregation node, perform the following steps:

1. Drag the Aggregation node on the left to the canvas and click the node to open the node configuration page.
2. Configure the Aggregation node by referring to **Table 7-20**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-20 Aggregation node configuration

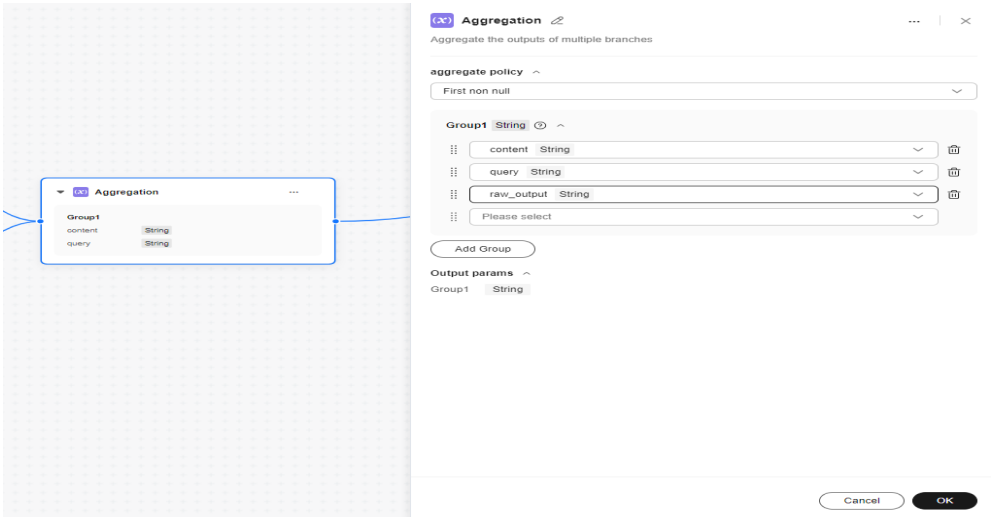
Configuration Type	Parameter Name	Description	Configuration Example (Scenario-based)
Parameter configuration	Output params	• **Param name**: The value is fixed to **Group1**. If there are multiple groups, set them as Group2, Group3, and so on, according to the number of groups. • **Param type**: depends on the variable data type of the corresponding aggregation group.	/
aggregate policy	-	You can specify a policy to aggregate all variables in each group. The aggregation policy is applied to variables in the same group. Currently, the aggregation policy can only be set to **First not null**. You can drag variables and adjust the variable positions. For example, if three variables output1, output2, and output3 are set in sequence in a group and are aggregated into one variable Group1, the value of output1 is used to assign a value to Group1 if output1 is not empty. If output1 is empty, the value of output2 is used, and so on.	Currently, the aggregation policy can only be set to **First not null**.
Aggregation group	-	By default, there is only one group Group1, which corresponds to one output variable Group1. All variables in the group are of the same type. If multiple variables need to be output, you can add multiple groups, which are named Group2, Group3, and so on.	/

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description	Configuration Example (Scenario-based)
Aggregation variable	-	Select the variables to be aggregated in the aggregation group. Each group can aggregate variables of only one data type. For example, multiple variables of the String type are aggregated into a String variable, and multiple variables of the Integer type are aggregated into an Integer variable.	For example, the Branch node or IntentDetection node splits a workflow into multiple branches. Each time the workflow is executed, the workflow determines to run a branch based on the branch condition or user intent. In this case, the output variables of other branches that are not executed are empty. You can connect multiple branches to the Aggregation node and set an output variable. No matter which branch is executed, the result can be referenced and accessed through the variable, avoiding repeated definition of the same semantic output variable by downstream nodes.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-61 Aggregation node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Aggregation node to other nodes.

7.4.2.4.13 Input Node

The Input node is used to provide input information during workflow running.

In some complex workflow scenarios, certain nodes need additional user input during execution. If the upstream node does not obtain the information, you can add an Input node to collect the information. The workflow execution is temporarily interrupted when the workflow reaches the Input node until the node collects necessary user input.

The Input node is an **optional** node. If it does not need to be configured, skip this section.

To configure the Input node, perform the following steps:

1. Drag the Input node on the left to the canvas and click the node to open the node configuration page.
2. Configure the Input node by referring to **Table 7-21**.

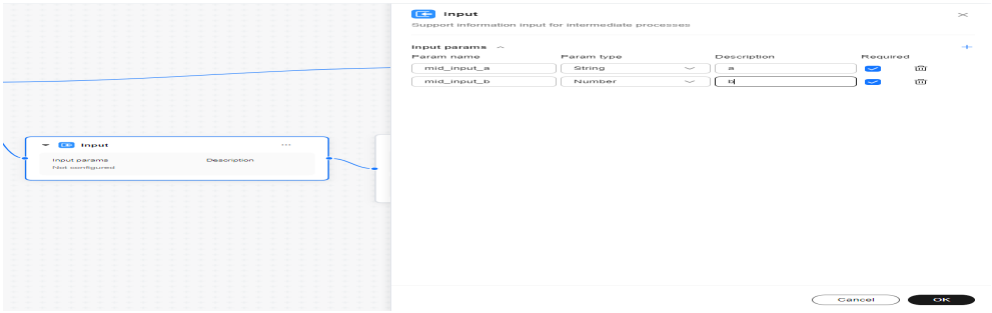
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-21 Input node configuration

Configuration Type	Parameter Name	Description	Configuration Example (Scenario-based)
Parameter configuration	Input params	You can configure one or more input parameters, which can be referenced by subsequent nodes. • **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Param type**: The value can be String, Integer, Number, Boolean, Object, Array, or File. • **Description**: description of the input parameter.	After one or more input parameters are configured, the workflow is interrupted when the workflow runs to the node. The workflow continues only when all input parameters are set and each input parameter can be referenced by the downstream node. For example, if a Plugin node needs an input parameter **city**, you can configure an Input node in the upstream node of the Plugin node to set the value of **city**, and then the Plugin node references the **city** parameter of the Input node.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-62 Input node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Input node to other nodes.

7.4.2.4.14 Workflow Node

The Workflow node is used to implement the nesting of workflows.

In a workflow, you can use another workflow as a step or node to encapsulate complex tasks. For example, you can encapsulate common and standardized task processing processes into different sub-workflows and call these sub-workflows in different branches of the main workflow to perform corresponding operations. Workflow nesting implements modular splitting and processing of complex tasks, making the workflow orchestration logic more flexible, clear, and easy to manage.

To configure the Workflow node, perform the following steps:

1. Drag the Workflow node on the left to the canvas and click the node to open the node configuration page.
2. Configure the Workflow node by referring to **Table 7-22**.

Table 7-22 Workflow node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	Input params	• The input structure of a workflow node depends on the input structure defined by the sub-workflow. The input structure cannot be customized. • You need to specify the data source for the input parameters in the Workflow node. The input parameters can be set to fixed values or reference the output parameters of upstream nodes.
	Output params	• The output structure of a workflow node depends on the output structure defined by the sub-workflow. The output structure cannot be customized. • **response_content** is a fixed output parameter of the workflow.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-63 Workflow node configuration example (sub-workflow)

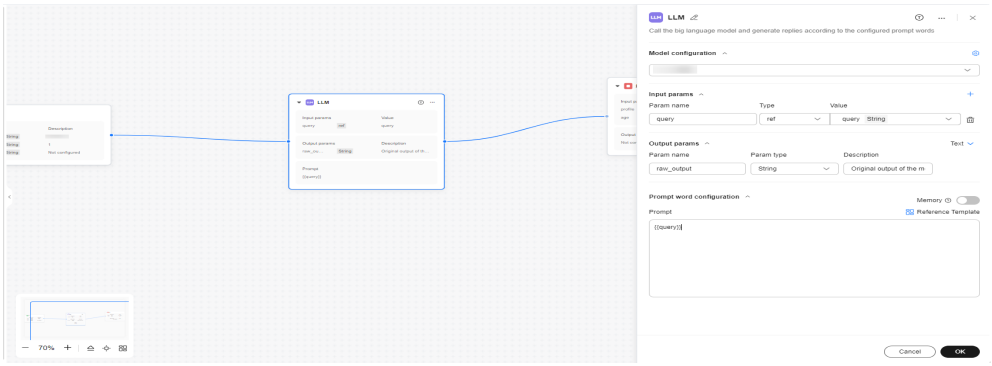
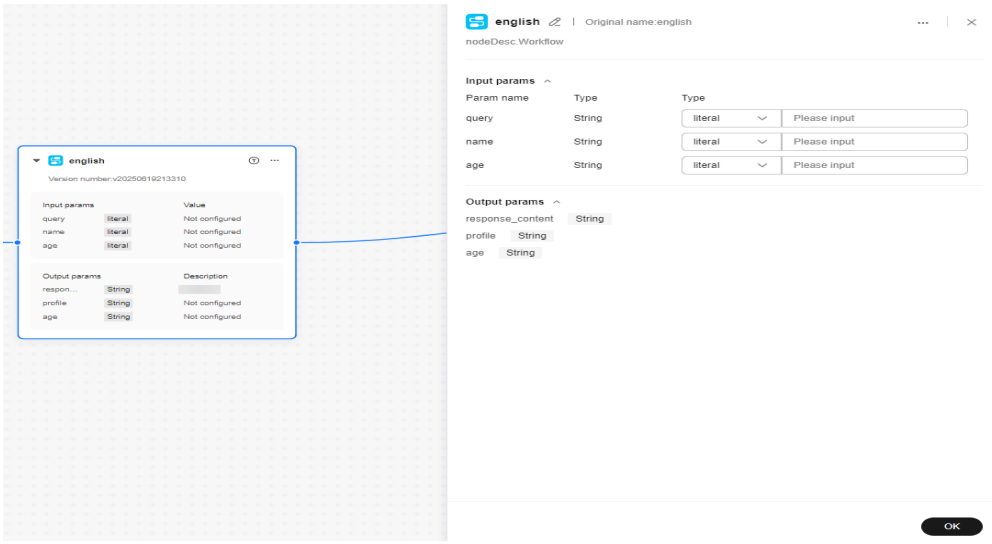


Figure 7-64 Workflow node configuration example - Workflow node (input and output correspond to the input and output of the sub-workflow)



3. After completing the configuration, click **OK**.
4. Connect the Workflow node to other nodes.

7.4.2.4.15 MCP Service Node

The MCP Service node is one of the core components for invoking third-party capabilities in workflows. As an important carrier for function extension, this node can invoke the MCP service to execute specific function tasks. Each MCP service is essentially a tool set. It can provide modular services to expand the capability boundary of workflows and complete more complex tasks.

MCP services are classified into preset services and personal services.

- Preset services: Multiple preset MCP services, such as AutoNavi Map, Ticket Query Tool, and Bing Search, are available on the platform. You can integrate and call them with one click. Developers can add preset MCP services to workflows or applications to enrich their capabilities.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Personal services: The platform allows developers to quickly create custom MCP services by configuring MCP service addresses for workflows or applications to call.

The MCP Service node is an **optional** node. If it does not need to be configured, skip this section.

To configure the MCP Service node, perform the following steps:

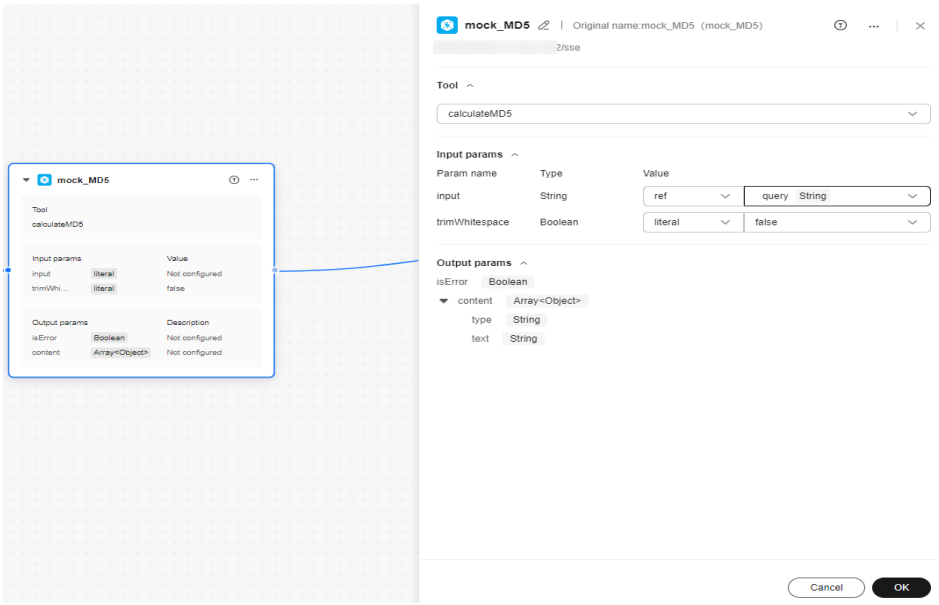
1. Drag the MCP Service node on the left to the canvas. On the **Personal Service** or **Pre installed services** tab page, click + to add an MCP service to the canvas. Some preset services cannot be directly added. You need to click **Enable Now** first. After the service is enabled, it can be added to the canvas.
   - The **Pre installed services** tab page lists the MCP services built in the platform.
   - Personal services are custom MCP services created by users. For details about how to create an MCP service, see **Creating a Custom MCP Service**.
2. Connect the MCP Service node to other nodes.
3. Click the added MCP Service node in the canvas and configure the node by referring to **Table 7-23**.

Table 7-23 MCP Service node configuration

Configuration Type	Parameter Name	Description
Parameter configuration	Tool	You can select a tool from the tool list of the current MCP service as the tool to be executed when the workflow runs to the node.
	Input params	• **Param name** and **Type**: imported from the plug-in metadata. You do not need to manually add a parameter. • **Type**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable value of a previous node that has been included in the workflow and the memory variable in the global configuration. This option applies to the scenario where plug-in input parameters need to be obtained from the output of a previous node. – **literal**: You can customize the value. This type applies to the scenario where the MCP service input parameter value is fixed.
	Output params	All output parameter information is automatically imported from the MCP service metadata. You do not need to manually configure the information.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-65 MCP Service node configuration example



4. After completing the configuration, click **OK**.

7.4.2.4.16 Agent Node

The Agent node provides the capability of using large models and calling large model tools. You can configure deployed models on the node. You can compile prompts and bind plug-ins to enable the model to process tasks.

The Agent node is an **optional** node. If it does not need to be configured, skip this step.

To configure the Agent node, perform the following steps:

1. Drag the Agent node on the left to the canvas and click the node to open the node configuration page.
2. Configure the Agent node by referring to **Table 7-24**.

Table 7-24 Agent node configuration

Configuration Type	Parameter Name	Description
Model configuration	Model selection	Select the model to be executed on this node. You can set parameters such as the generation diversity of the model on this node to make the model effectiveness meet your expectation.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

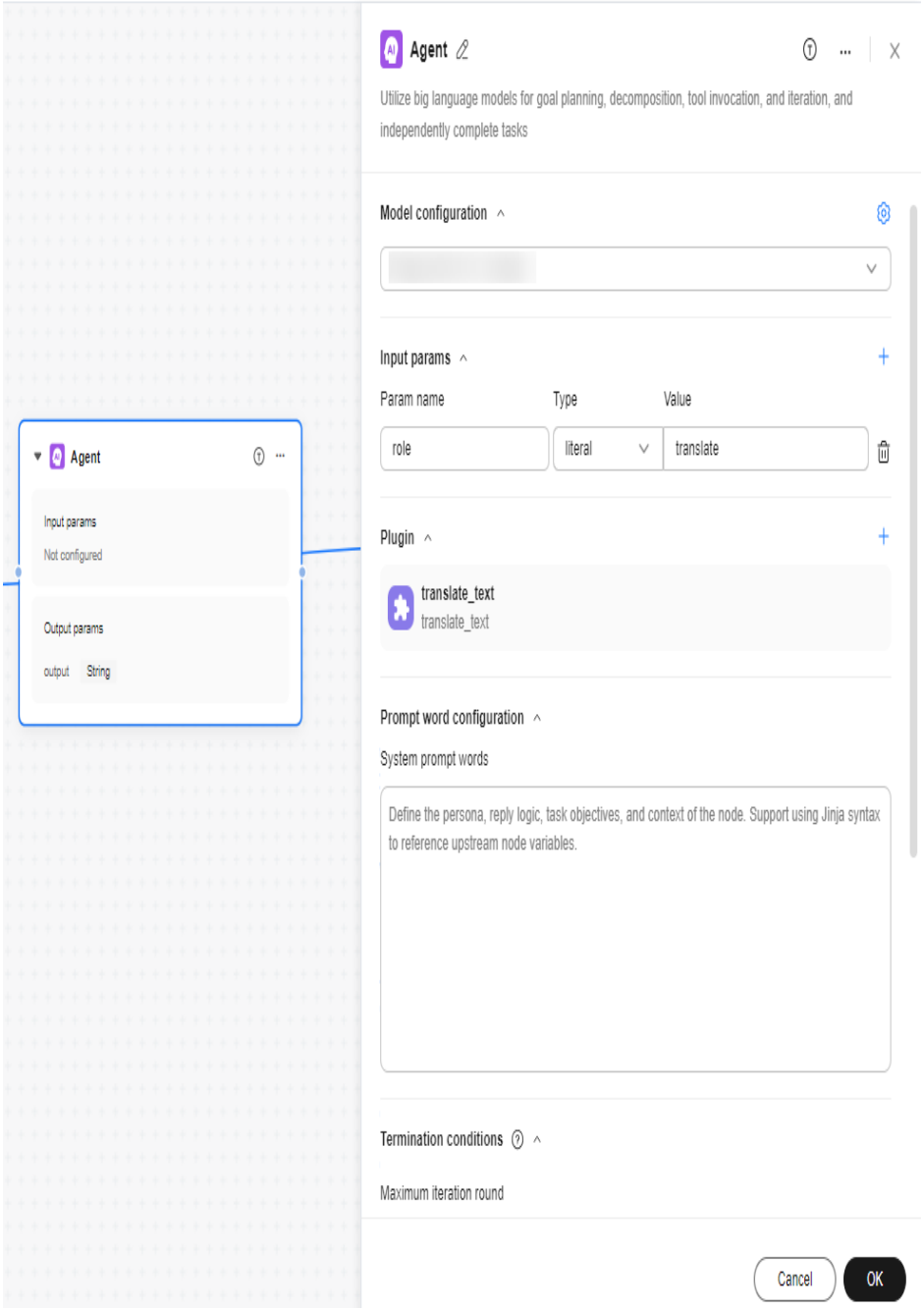
Configuration Type	Parameter Name	Description
	Top P	During output, the model selects words with the highest probability until the total probability of these words reaches the Top P value. The Top P value can restrict the model to select these high-probability words, thereby controlling the diversity of output content. You are advised not to adjust this parameter together with the **Temperature** parameter.
	Temperature	Controls the randomness of the generation result. A higher temperature makes the model output more diverse and innovative. A lower temperature makes the output more compliant with the instruction requirements, but reduces the diversity of the model output.
Parameter configuration	Input params	• **Param name**: The value can contain only letters, digits, and underscores (_), and cannot start with a digit. • **Type** and **Value**: **Type** can be set to **ref** and **literal**. – **ref**: You can select the output variable values of the previous nodes contained in the workflow and the memory variables in the global configuration. – **literal**: The value can be customized.
	Plugin	You can bind a manually created plug-in or a preset plugin. When the model detects that a tool needs to be called to complete a task, the model extracts parameters based on user input to call the plug-in and summarizes the plug-in execution result.
	System prompt words	System prompts instruct the model to complete tasks better. When configuring prompts, you can use the {{variable}} format to reference the parameters defined in the input parameters of the current node. The replaced content is transferred to the model.
	Output params	Output parameters are the outputs of the last round on the Agent node.
Termination Conditions	Maximum iteration round	This parameter specifies the maximum number of interaction rounds with the model. If no parameter is extracted when the maximum number of interaction rounds is reached, the Agent node exits.
	Plugin execution successful	After this function is enabled, the plug-in can be bound to the agent. After any selected plugin is successfully executed, the Agent node can be exited.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Configuration Type	Parameter Name	Description
	Identified that the user has an intention to exit	After this function is enabled, once the LLM recognizes that the user has the intention to exit, the Agent node can be exited.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-66 Agent node configuration example



3. After completing the configuration, click **OK**.
4. Connect the Agent node to other nodes.

7.4.3 Using APIs to Call a Workflow

After the trial run of a workflow is successful, you can call the workflow through APIs.

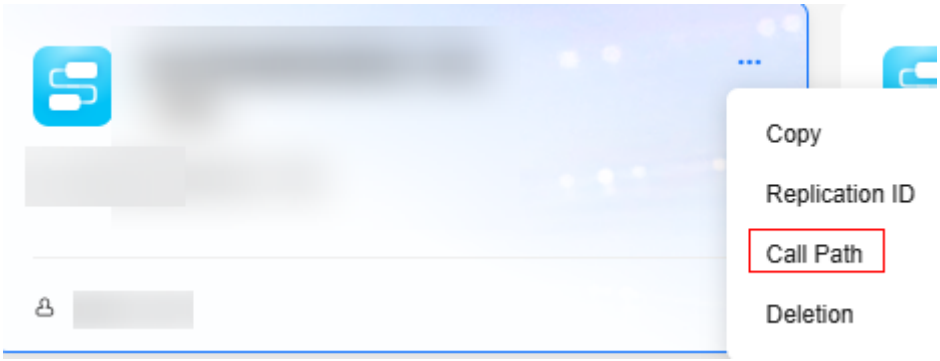
Obtaining the Call Path

To obtain the call path of a workflow, perform the following steps:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

1. Log in to ModelArts Studio and choose **Agent Dev** in the navigation pane.
2. On the **Workbench** page, click the **Workflows** tab. Click ··· and choose **Call Path** of the required workflow.

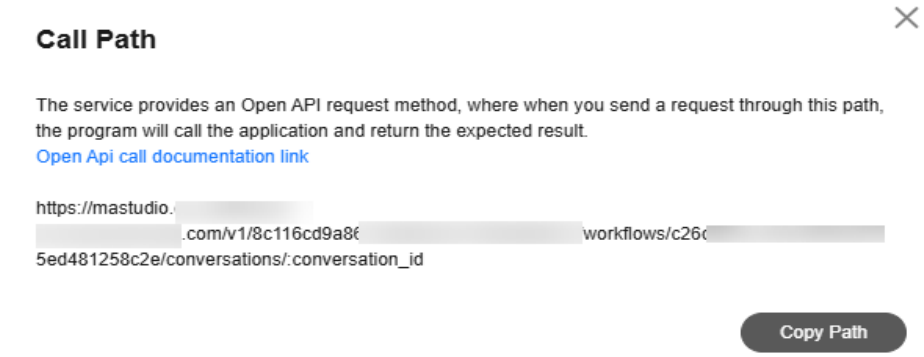
   Figure 7-67 Obtaining the call path of a workflow (1)

3. In the **Call Path** dialog box, click **Copy Path**.

   The **conversation_id** parameter indicates the session ID, which uniquely identifies a session. You can set the session ID to any value in the standard UUID format. For example, **conversation_id** in the following figure can be replaced with 550e8400-e29b-41d4-a716-446655440000.

   Figure 7-68 Obtaining the call path of a workflow (2)

Obtaining a Token

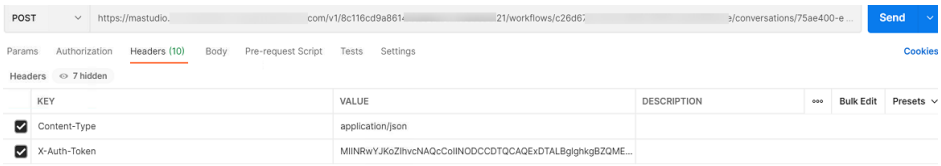
Obtain the token by following the instructions provided in section "Calling REST APIs" > "Authentication" in *API Reference*.

Using Postman to Call APIs

1. Create a POST request in Postman and enter the workflow call path. For details, see **Obtaining the Call Path**.
2. Set the request header parameters.
   - **KEY**: **Content-Type**; **VALUE**: **application/json**
   - The header parameter **X-Auth-Token** is mandatory, and the parameter value is the token obtained in **Obtaining a Token**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-69 Configuring the request header



3. In Postman, choose **Body** > **raw**. The following is an example of configuring the request body:

   The **inputs** parameter is a question raised by the user, and is used as an input of the workflow. If the workflow has a plug-in, add the **plugin_configs** parameter. The **plugin_id** parameter indicates the plug-in ID. For details about how to obtain the plug-in ID, see **Managing Plug-ins**.

   Request with plug-ins

   ```
   {
     "inputs": {
       "query": "Hello"
     },
     "plugin_configs": [
       {
         "plugin_id": "xxxxxxxxx",
         "config": {
           "key": "value"
         }
       }
     ]
   }
   ```

   Request without a plug-in

   ```
   {
     "inputs": {
       "query": "Introduce Beijing."
     }
   }
   ```

4. Click **Send** on Postman to send the request. If status code 200 is returned, the API is successfully called.

   The following is an example of a workflow with a Questioner node:

   ```
   {
       "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
       "messages": [
           {
               "role": "assistant",
               "content": "Hello"
           }
       ],
       "status": {
           "code": 1,
           "desc": "succeeded"
       },
       "start_time": 1734336269313,
       "end_time": 1734336270908
   }
   ```

   The following is an example of a workflow without a Questioner node:

   ```
   {
       "conversation_id": "2c90493f-803d-431d-a197-57543d414317",
       "outputs": {
           "responseContent": "Hello. How can I help you?"
       },
       "messages": [],
       "status": {
           "code": 1,
           "desc": "succeeded"
       },
       "start_time": 1734337068533,
   ```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
    "end_time": 1734337082545
}
```

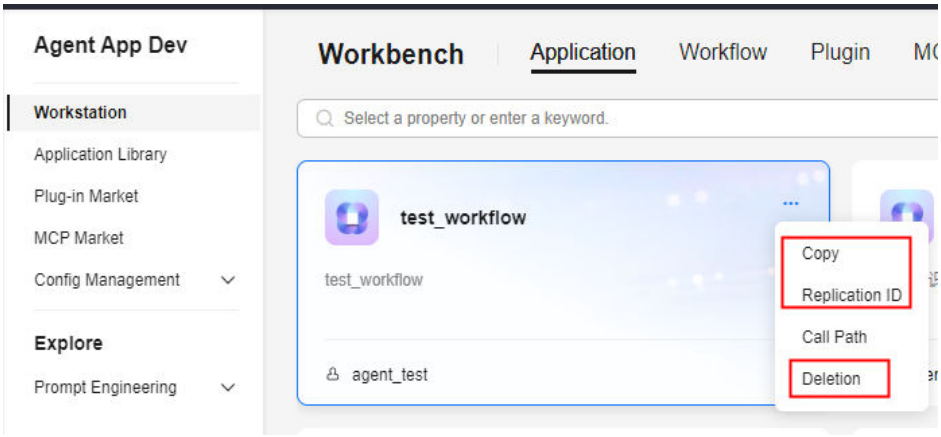
7.4.4 Managing Workflows

On the agent development platform, you can copy and obtain workflow IDs, as well as delete, import, and export workflows.

Obtaining a Workflow ID and Deleting a Workflow

1. Log in to the agent development platform.
2. Choose **Workbench** > **Workflow**.
3. Click ··· > **Copy** to copy the current workflow.
4. Click ··· > **Copy ID** to obtain the current workflow ID.
5. Click ··· > **Delete** to delete the current workflow.

Figure 7-70 Obtaining a Workflow ID and Deleting a Workflow



NOTE

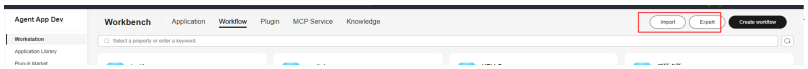
Deletion is a high-risk operation. Before deleting a workflow, ensure that the workflow is no longer used.

Exporting and Importing Workflows

The platform supports workflow export and import. When a workflow is exported, the configurations such as plug-ins associated with the workflow are also exported.

1. Log in to the agent development platform.
2. Choose **Workbench** > **Workflow**.
3. Export a workflow.
   a. Click **Export** in the upper right corner of the page.

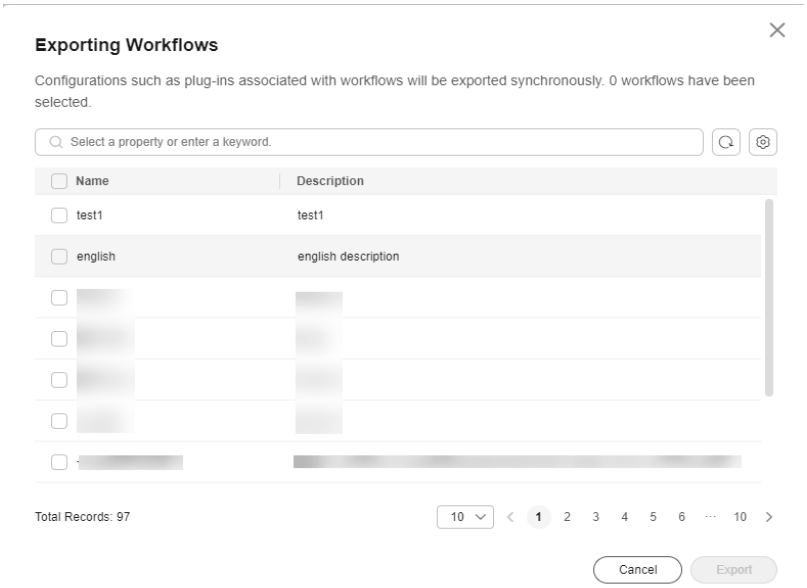
Figure 7-71 Export



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

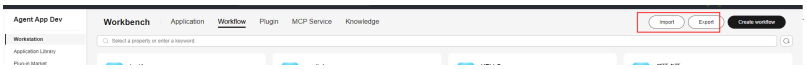
b. In the **Export Workflow** dialog box, select a workflow and click **Export**. The workflow is downloaded to the local host as a JSONL file.

Figure 7-72 Export Workflow



4. Import a workflow.
   a. Click **Import** in the upper right corner of the page.

Figure 7-73 Import

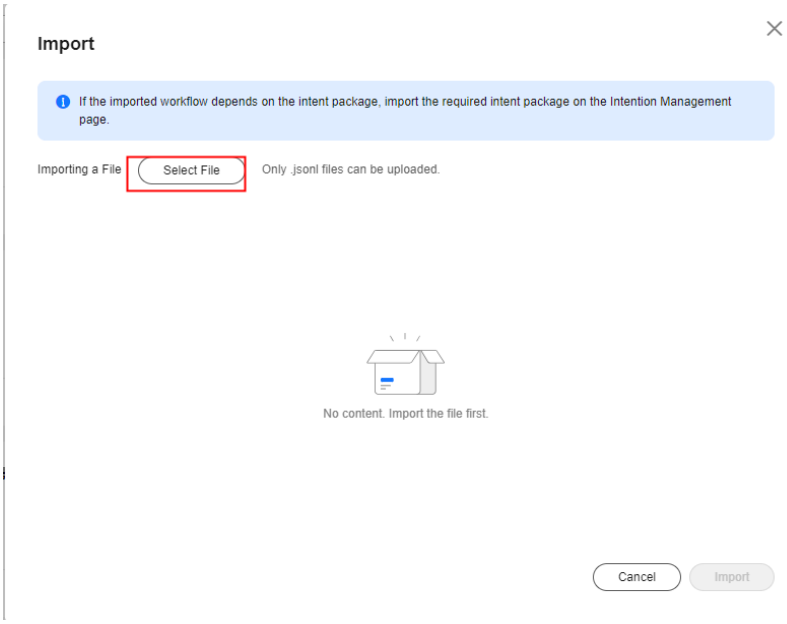


   b. In the **Import** dialog box, click **Select File** and select the JSONL file to be imported.
   c. Select the content to be parsed.

      The platform automatically parses the JSONL file. If the file to be parsed already exists on the platform, this file will automatically overwrite the existing file on the platform.
   d. Click **Import**. The imported workflow is displayed on the **Workflow** tab page of the **Workbench** page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-74 Import



7.5 Plug-in Development and Configuration

7.5.1 Introduction to Plug-ins

On the agent development platform, plug-ins are important extensions of large model capabilities. In modular mode, plug-ins can provide more professional skills and complex task processing capabilities for large models, meeting user requirements more efficiently in diverse scenarios.

Applications interconnected with plug-ins can provide capabilities that are not supported by the large model. Plug-ins provide various external service APIs. When a task is executed, the model detects the applicable plug-ins based on the prompts, automatically invokes the plug-ins, obtains the result from the external service, and returns the result. This design enables the Agent to intelligently process complex tasks and even solve problems across domains, implementing automatic processing of complex problems.

The agent development platform supports two types of plug-ins:

- Preset plug-ins: These plug-ins are provided by the platform for developers and users. They are ready to use and do not require custom development. For example, the Python interpreter plug-in provided by the platform can automatically generate Python code based on the questions entered by users, and execute the code to obtain the result. This plug-in provides powerful computing, data processing, and analysis functions for agents. You only need to add the plug-in to an application to extend its functions.
- Custom plug-ins: To meet more personalized requirements, the platform allows developers to create custom plug-ins. APIs can be quickly created as plug-ins through configuration and called by agents. In this way, developers can add dedicated functions to applications based on specific requirements.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

7.5.2 Creating a Plug-in

7.5.2.1 Creating a Plug-in Based on an API

Creating a Plug-in

To create a plug-in, perform the following steps:

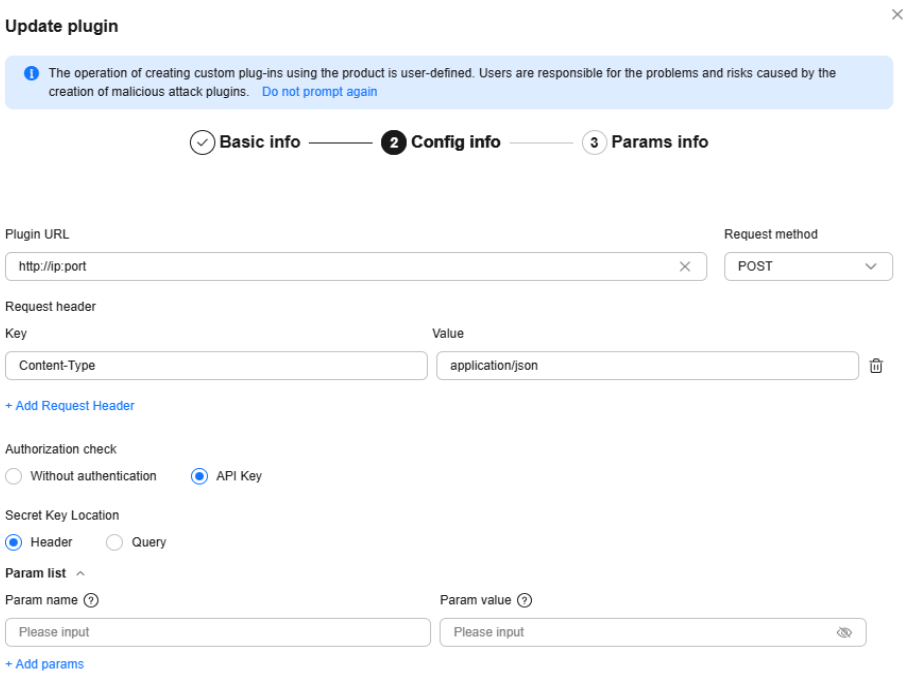
1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. On the **Plugin** tab page, click **Create plugin** in the upper right corner.
3. In the **Create plugin** dialog box, enter the plugin name and description, and set **Only I Can See** based on the actual requirement. (After it is enabled, the plugin is visible only to the creator.) Click the image to upload the plug-in icon and click **Next**.
4. In the **Config Info** step, set the parameters based on **Table 7-25**.

Table 7-25 Plug-in information configuration

Parameter Name	Description
Plugin URL	Request URL of the plug-in service. • The URL protocol can only be HTTP or HTTPS. • The system checks whether the URL is in the standard format. • By default, the IP address corresponding to the URL cannot be an intranet IP address. Otherwise, the registration fails.
Request Method	Request method of the plug-in service. The value can be POST or GET.
Authorization check	Whether an API can be called only after being authenticated. • **Without authentication**: APIs can be accessed publicly without any form of authentication or authorization. • **API Key**: A unique API key is provided for authentication when an API is called. Enter the key location, that is, **Header** or **Query**. Set the key authentication parameter names and key value.
Header	Enter the request header information of the API. For example: • **Key**: Content-Type • **Value**: application/json

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-75 API request configuration example



NOTE

Using HTTP for custom plug-ins or selecting the **Without authentication** mode can introduce security risks.

5. Click **Next step**. In the **Params Info** step, set parameters based on **Table 7-26**.

Table 7-26 Plug-in parameter configuration

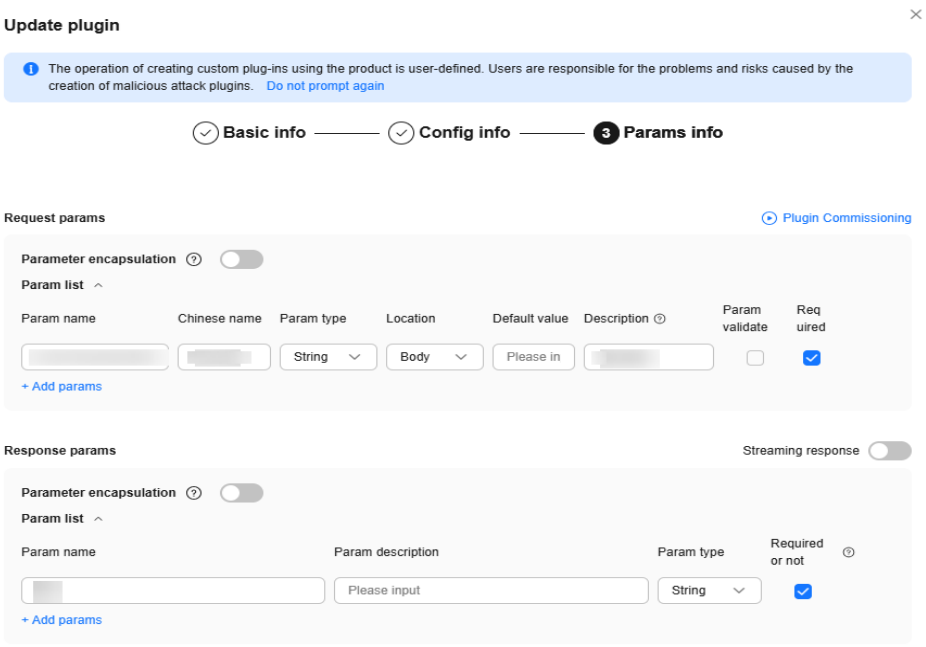
Category	Parameter Name	Description
Request params	Parameter encapsulation	After this function is enabled, request parameters are encapsulated into a list (array) structure. Plug-in APIs whose input parameters are in array format can be adapted. For example, the original parameter list is {"a":"string", "b":1}, and the parameter list after encapsulation is enabled is [{"a":"string", "b":1}].
	Parameter Name	Name of a parameter. The parameter name is used by the large model to parse the parameter meaning.
	Chinese name	Chinese name of the parameter.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Category	Parameter Name	Description
	Parameter Type	Data type of the parameter value. The options are **String**, **Integer**, **Number**, etc.
	Secret Key Location	Location of the current parameter in the request. The value can be **Body**, **Headers**, or **Query**.
	Default value	Default parameter value.
	Description	Parameter description. The meaning and requirements of a parameter must be described as accurately as possible so that the Agent can accurately extract parameters.
	Param validate	You can set the validation rule for the current parameter.
	Required or not	Whether the parameter is required.
Response params	Parameter encapsulation	After this function is enabled, request parameters are encapsulated into a list (array) structure. Plug-in APIs whose input parameters are in array format can be adapted. For example, the original parameter list is {"a":"string", "b":1}, and the parameter list after encapsulation is enabled is [{"a":"string", "b":1}].
	Parameter Name	Name of a response parameter. The parameter name is used by the large model to parse its outputs.
	Explanation	Name of a response parameter. The parameter description is used by the large model to parse its outputs.
	Parameter Type	Data type of the parameter value. The options are **String**, **Integer**, **Number**, etc.
	Required or not	If this function is enabled, this parameter must be extracted. If this function is disabled, this parameter can be left empty or the default value is used.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-76 Configuring API request and response parameters



7.5.2.2 Importing Plug-ins Using JSON Files

You can import plug-ins using JSON files.

1. Log in to the agent development platform.
2. Choose **Workbench** > **Plugin**.
3. Import a plug-in.
   a. Click **Import** in the upper right corner of the page.
   b. On the **Import** page, click **Select File** and select the JSONL file to be imported.
   c. Select the content to be parsed.

      The platform automatically parses the JSONL file. If the file to be parsed already exists on the platform, this file will automatically overwrite the existing file on the platform.
   d. Click **Import**. The imported plug-in is displayed on the **Plugin** tab page of the **Workbench** page.

7.5.3 Managing Plug-ins

On the agent development platform, you can obtain plug-in IDs, delete, import, and export plug-ins.

Deleting a Plug-in Based on the Plug-in ID

1. Log in to the agent development platform.
2. Choose **Workbench** > **Plugin**.
3. Click ··· and choose **Copy ID** to obtain the plug-in ID.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4. Click ··· and choose **Delete** to delete the current plug-in.

   NOTE

   Deleting a plug-in is a high-risk operation. Before deleting a plug-in, ensure that it is no longer used.

Exporting and Importing Plug-ins

The platform supports plug-in export and import.

1. Log in to the agent development platform.
2. Choose **Workbench** > **Plugin**.
3. Export a plug-in.
   a. Click **Export** in the upper right corner of the page.
   b. In the **Export Plugin** dialog box, select a workflow and click **Export**. The plug-in is downloaded as a JSONL file to the local PC.
4. Import the plug-in.
   a. Click **Import** in the upper right corner of the page.
   b. In the **Import** dialog box, click **Select File** and select the JSONL file to be imported.
   c. Select the content to be parsed.

      The platform automatically parses the JSONL file. If the file to be parsed already exists on the platform, this file will automatically overwrite the existing file on the platform.
   d. Click **Import**. The imported plug-in is displayed on the **Plugin** tab page of the **Workbench** page.

7.6 Knowledge Base Development and Configuration

7.6.1 Knowledge Base Introduction

The knowledge base function of the platform is used to store data such as text documents in vectorized mode and retrieve knowledge, and provides retrieval enhancement capabilities for applications and workflows. You can easily import data into the knowledge base, whether it is a text file, presentation, or spreadsheet file, without additional conversion or formatting.

The knowledge base supports the import of local documents in the following formats:

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Table 7-27 Supported document formats

Document Type	Constraint	Size Requirements
Knowledge document	You can upload documents in common text formats, including .doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, and .md.	• The maximum size of a single document that can be uploaded is 10 MB.
FAQ	Text can be uploaded based on a template file in Word or Excel format. For details about the template file, see **Excel Template Format** and **Word Template Format**.	• The maximum size of an Excel file is 10 MB, and the maximum number of data records is 10,000. The file cannot contain empty lines. Data after an empty line will be ignored. • The maximum size of a Word file is 128 MB. The file must be filled according to the template format. [QUESTION] is the question, and the content is the answer. Images in the document must be in PNG, JPG, or JPEG format.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Excel Template Format

Table 7-28 Excel template example

Answer	Question	Similar Question 1 (Optional)	Similar Question 2 (Optional)	Similar Question 3 (Optional)	Similar Question 4 (Optional)	Directory (Optional)	Tag (Optional, Separated by Commas)
Cloud Search Service (CSS) is a fully hosted distributed search service based on Elasticsearch. You can use it for structured and unstructured data search, and use AI vectors for combine search, statistics, and reports.	What is Cloud Search Service?	What is Huawei Cloud CSS?	What is CSS?	What is Huawei Cloud's Cloud Search Service?	What is the Cloud Search Service from Huawei Cloud?	-	-
CSS can be used to build search boxes for websites and apps to improve user experience. You can also build a log analysis platform with it, facilitating data-driven O&M and business operations. CSS vector search can help you quickly build smart applications, such as AI-based image search, recommendation, and semantic search.	What are the use cases of CSS?	What are the use cases of Huawei Cloud CSS?	What is CSS used for?	What are the functions of CSS?	What do I do with CSS?	-	-

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Answer	Question	Similar Question 1 (Optional)	Similar Question 2 (Optional)	Similar Question 3 (Optional)	Similar Question 4 (Optional)	Directory (Optional, Separated by Commas)	Tag (Optional, Separated by Commas)
You can access SMS from the console, or using HTTPS-based APIs.	How do I access CSS?	What access methods does CSS support?	How do I access Huawei Cloud's Cloud Search Service?	How do I use CSS?	-	-	-

Word Template Format

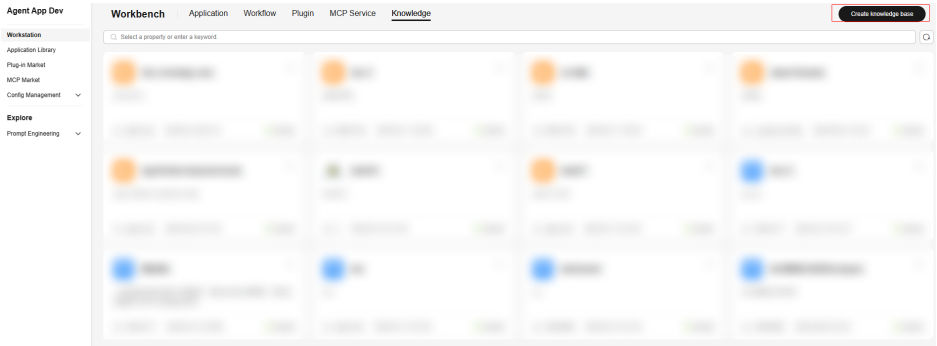
The knowledge base template in Word format writes all knowledge into a Word document in the form of Q&A pairs. Each Q&A corresponds to a title pair. A question starts with [QUESTION] and the text attribute is set to Heading 1. The answer is in a new line and the text attribute is set to Body.

7.6.2 Creating a Knowledge Base and Uploading Documents

Creating a Knowledge Base

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. On the **Knowledge** tab page, click **Create knowledge base** in the upper right corner.

Figure 7-77 Creating a knowledge base



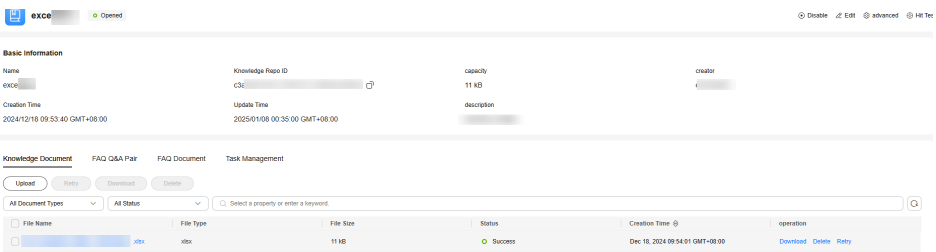
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3. In the **Create knowledge base** dialog box, set the basic information, Embedding model, Rerank model, and document parsing and splitting policies.
   - **Basic Information**: Set the knowledge base icon, name, and description.
   - **Model Configuration**: Select the Embedding model and Rerank model.
   - **Parsing Configuration**: Configure document parsing.
   - **Split Configuration**: Configure document splitting.
4. After the configuration is complete, click **OK**.

Uploading Documents

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. On the **Knowledge** tab page, click on a knowledge base card to go to the details page.
3. Select the type of the document to be uploaded.
   - Knowledge Document
     i. On the **Knowledge Document** tab page of the details page, click **Upload**.
     ii. In the displayed dialog box, click the **click here to upload** link to upload a local file to the knowledge base.
     iii. Upload a file in any of the following formats: .doc, .docx, .pdf, .pptx, .ppt, .xlsx, .xls, .csv, .wps, .png, .jpg, .jpeg, .bmp, .gif, .tiff, .tif, .webp, .pcx, .ico, .psd, .dps, .et, .txt, .ofd, or .md.
     iv. The size of a single file cannot exceed 10 MB.
     v. A maximum of 300 files can be uploaded at a time.
   - FAQ Document
     i. On the **FAQ Document** tab page of the details page, click **Upload**.
     ii. In the displayed dialog box, click the **click here to upload** link to upload a local file to the knowledge base. The file must be uploaded based on the template. You can download a template.
     iii. The maximum size of an Excel file is 10 MB, and the maximum number of data records is 10,000. The file cannot contain empty lines. Data after an empty line will be ignored.
     iv. The maximum size of a Word file is 10 MB.
     v. A maximum of 300 files can be uploaded at a time.

Figure 7-78 Uploading documents



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

NOTE

After a knowledge base is created, you can click the knowledge base to go to the details page, click the corresponding tab, and upload documents.

You can modify the configured document parsing and splitting policies by clicking **advanced** in the upper right corner of the knowledge base details page. After the modification, you can parse and split the document again.

7.6.3 Knowledge Base Hit Test

The platform supports the hit test on the created knowledge base to evaluate the effectiveness and accuracy of the knowledge base.

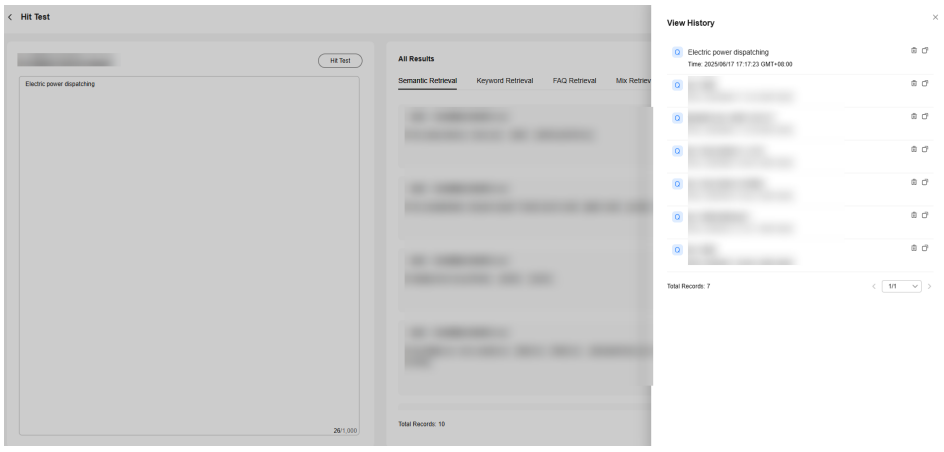
The hit test matches the user's query with the content in the knowledge base, provides the related information, and sorts the information based on the matching degree.

The knowledge base hit test procedure is as follows:

1. Log in to the agent development platform.
2. On the **Knowledge** tab page of the **Workbench** page, click the required knowledge base, and click **Hit Test** in the upper right corner.
3. Enter a question in the text box and click **Hit Test**. The lower part of the page displays multiple matched contents according to different search modes and sorts the content in descending order by matching score.

   You can evaluate whether the current knowledge base meets the requirements based on the score and the amount of matched information.
4. Click **View History** to view the historical questions entered by the user.

Figure 7-79 Viewing history



7.6.4 Managing Knowledge Bases

Adding and Deleting Documents in the Knowledge Base

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. Click the **Knowledge** tab.
3. Click the required knowledge base. The details page is displayed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- Add knowledge documents to the knowledge base. Click **Continue Upload** in the upper right corner to upload local documents to the current knowledge base.
- Delete a knowledge document from the knowledge base. On the **Knowledge Document** tab page, click **Delete** in the **Operation** column to delete the current knowledge document.

Enabling or Disabling a Knowledge Base

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. Click the **Knowledge** tab. Click the icon in the upper right corner of a knowledge base card to view the knowledge base status. Click **Enable** or **Disable** to enable or disable the knowledge base.
3. Click the required knowledge base. The details page is displayed. Click **Enable** or **Disable** in the upper right corner.

NOTE

A knowledge base can be referenced in applications and workflows only when it is enabled.

If a knowledge base that has been referenced by applications or workflows is disabled, no search result is returned. Exercise caution when performing this operation.

Obtaining a Knowledge Base ID and Deleting a Knowledge Base

1. Log in to the agent development platform.
2. In the navigation pane, choose **Workstation**. Click the **Knowledge** tab.

- Click ··· > **Copy ID** to obtain the current knowledge base ID.
- Click ··· > **Delete** to delete the current knowledge base.

NOTE

Deleting a knowledge base is a high-risk operation. Before deleting a knowledge base, ensure that the knowledge base is no longer used.

7.7 MCP Service Development and Configuration

7.7.1 MCP Service Introduction

Model Context Protocol (MCP) is an open standard introduced by Anthropic. It is designed to unify the communication protocols between large language models (LLMs) and external data sources and tools.

MCP aims to address the issue of AI models failing to reach their full potential due to data silos. It enables AI applications to securely access and operate on both local and remote data, providing an interface for seamless connectivity with all resources.

7.7.2 Creating an MCP Service

On the console, you can choose **Agent App Dev** > **MCP Market** to use preset MCP services, or choose **Agent App Dev** > **Workstation** and click **MCP Service** to

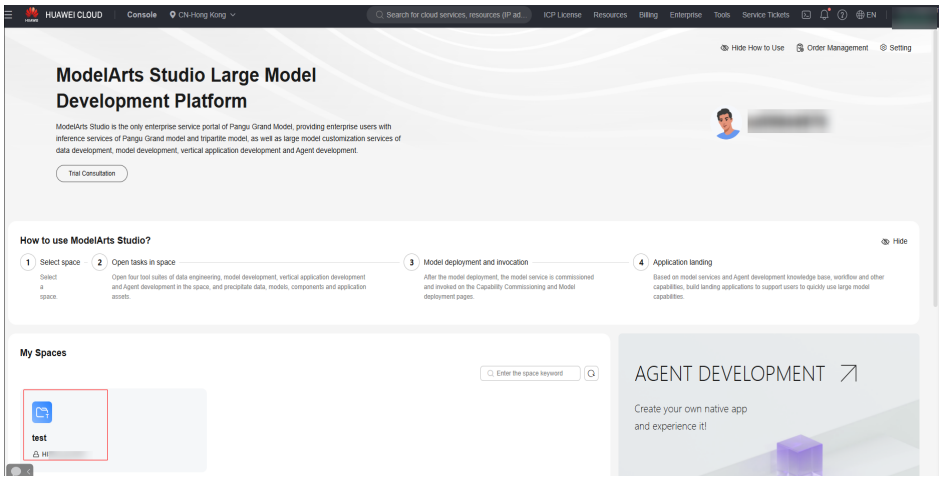
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

create an MCP service. ModelArts Studio supports one-click adding of MCP services during intelligent agent and workflow application development, enhancing the capability of calling agent tools.

Using a Preset MCP Service

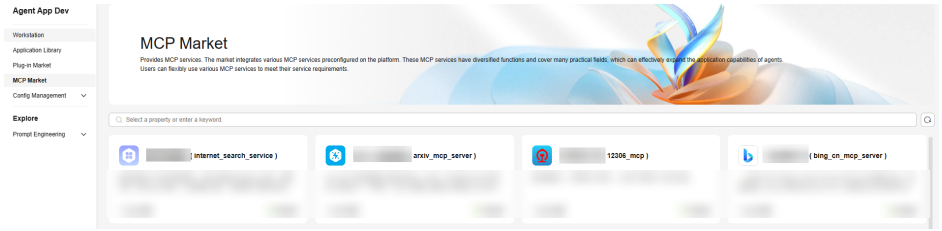
1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 7-80 My Spaces

2. In the navigation pane, choose **Agent Dev**.
3. Choose **MCP Market**.

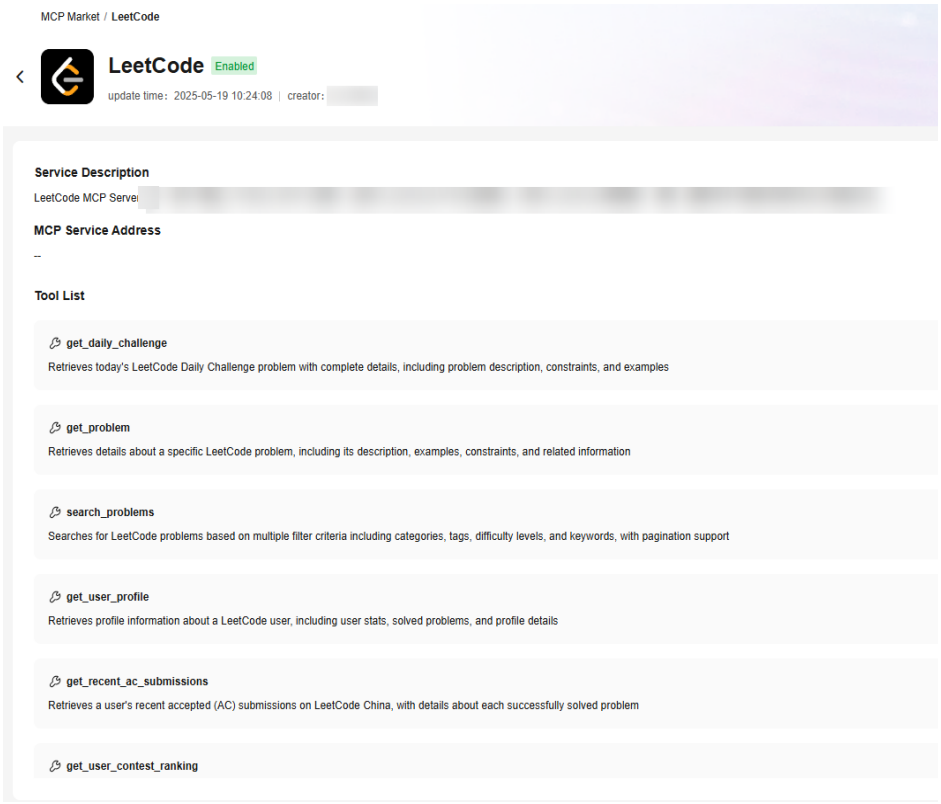
   Figure 7-81 MCP Market

4. Search for the target service by attribute type or search for the MCP service name. Click the name of the target service to access its details page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-82 Service details page



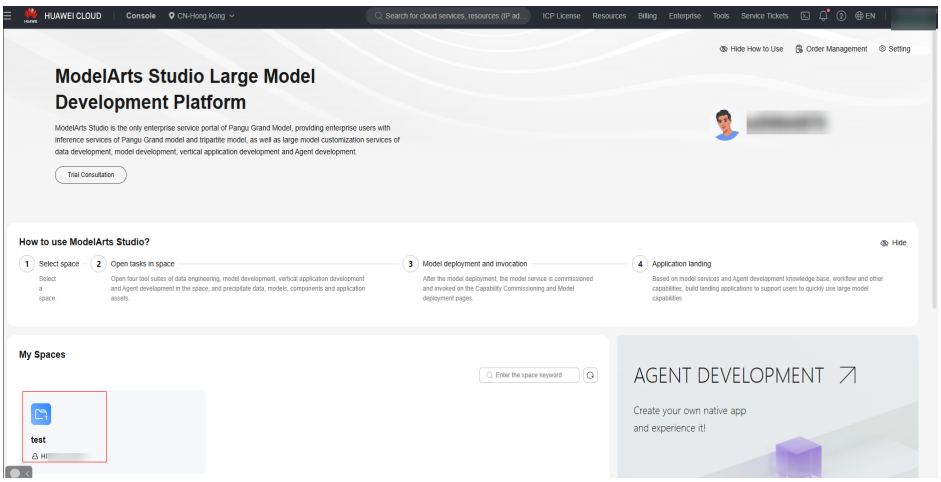
5. Perform operations as prompted. If the MCP service requires API key authentication, enter the API key and click **OK** to enable the MCP service.

Creating a Custom MCP Service

If you already have tools and API resources available, you can use the MCP server to configure your MCP extension. After the MCP server is built, you only need to configure the corresponding tool on the agent development platform to create your own MCP service.

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 7-83 My Spaces



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

2. In the navigation pane, choose **Agent Dev**.
3. Click the **MCP Service** tab and click **Create MCP Service**.
4. In the **Create MCP Service** dialog box, set **MCP Service Name**, **English Name**, **MCP Service Address**, **Description**, and **Authorization check** by referring to **Table 7-29**, and click **Test and Next**.

Table 7-29 Parameters for configuring MCP service information

Parameter	Description
MCP Service Name	Chinese name of an MCP service, which is displayed externally.
English Name	English name of an MCP service, which uniquely identifies the MCP service in internal interaction. It is used in API calling, inter-system communication, and code reference. Ensure that the value is unique under the same tenant.
MCP Service Address	SSE address of the MCP service. • The service URL supports only HTTP and HTTPS. • The system checks whether the URL is in the standard format. • By default, the IP address corresponding to the URL cannot be an intranet IP address. Otherwise, the registration fails. **NOTE** Public MCP services on the website and **user-created MCP services** are supported.
Description	MCP service description.
Only I can see	Whether to be visible only to the creator.
Authorization check	Whether authentication is required when the MCP service is called. • **Without authentication**: MCP services can be accessed publicly without any form of authentication or authorization. • **API Key**: A unique API key is provided for authentication when an MCP service is called. Set **Secret Key Location** to determine whether the secret information is contained in the header or query. Set the API key authentication parameter names and values. **NOTE** If the service address contains authentication information, you are advised to use the authorization check (API Key) function of the platform.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-84 Creating an MCP Service

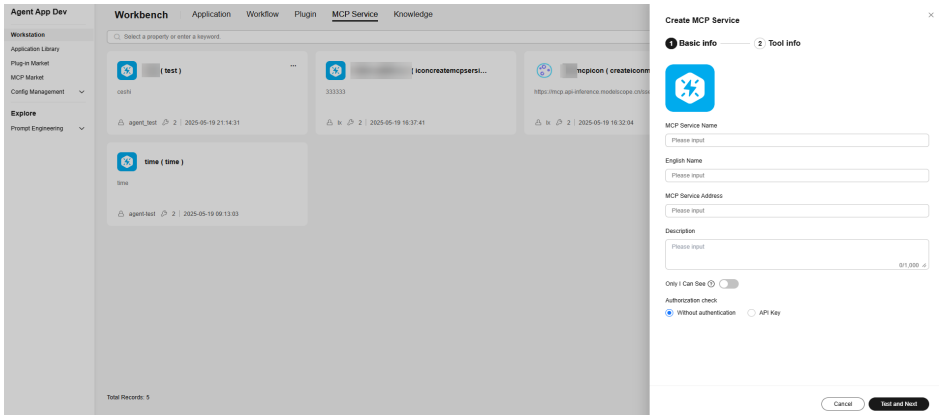
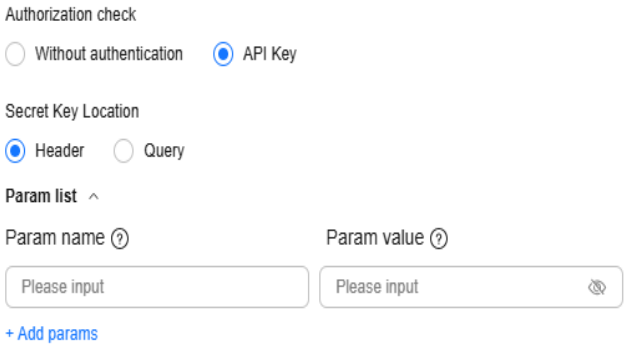
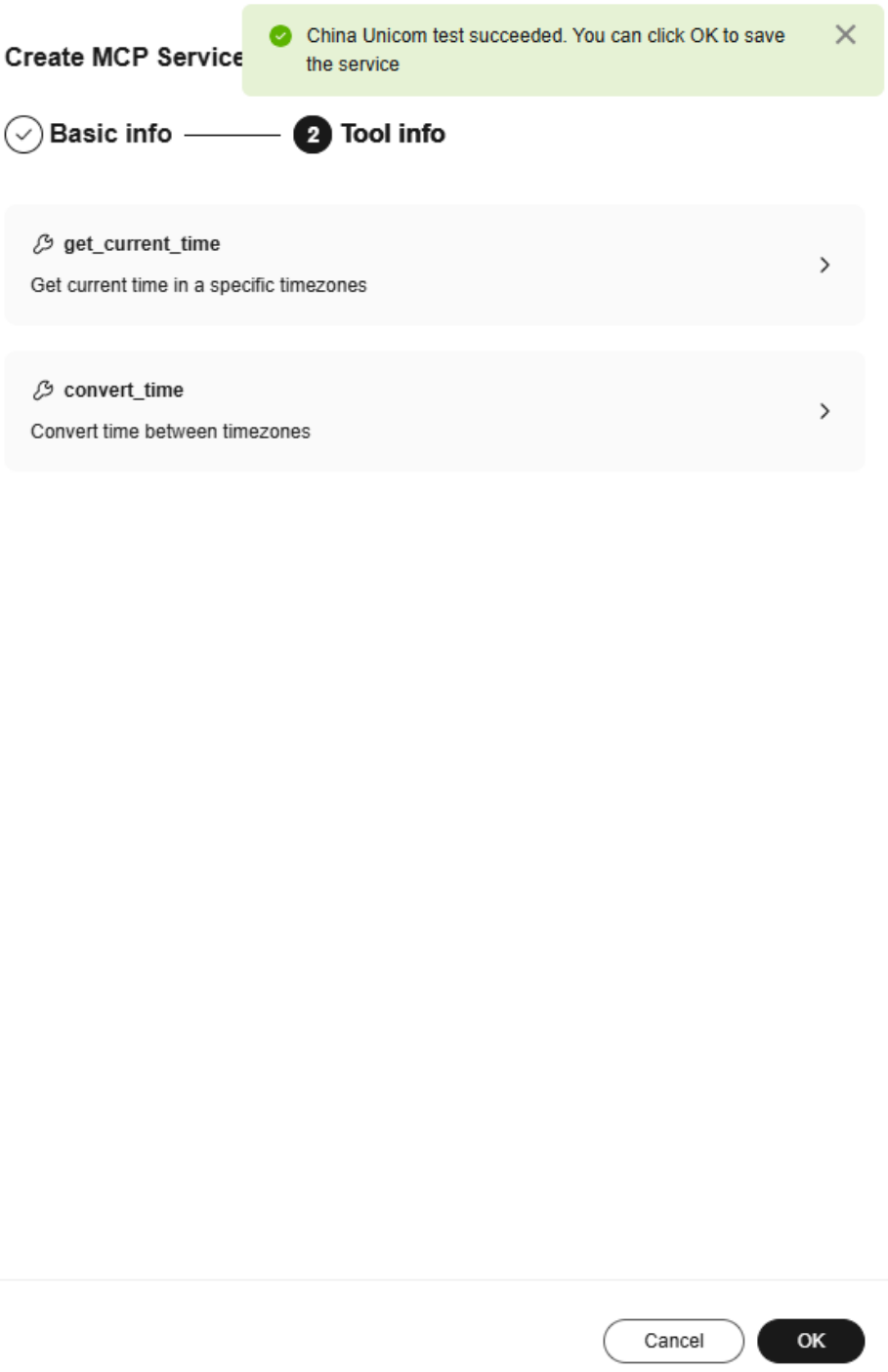


Figure 7-85 Authorization check example



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 7-86 Connectivity test succeeded



5. After the test is complete, click **OK**.

Using the MCP Service

On the agent development platform, developers can add preset and custom MCP services when developing agent and workflow applications.

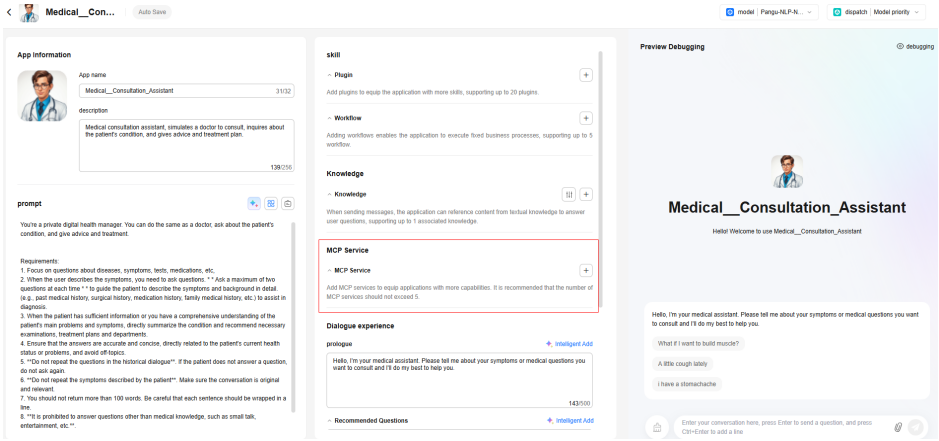
- **Configure the MCP service during agent development.**

  Step 1: Log in to the agent development platform and choose **Workstation** from the navigation pane.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Step 2: Click the **Application** tab page. On the page that is displayed, click **Create App**.

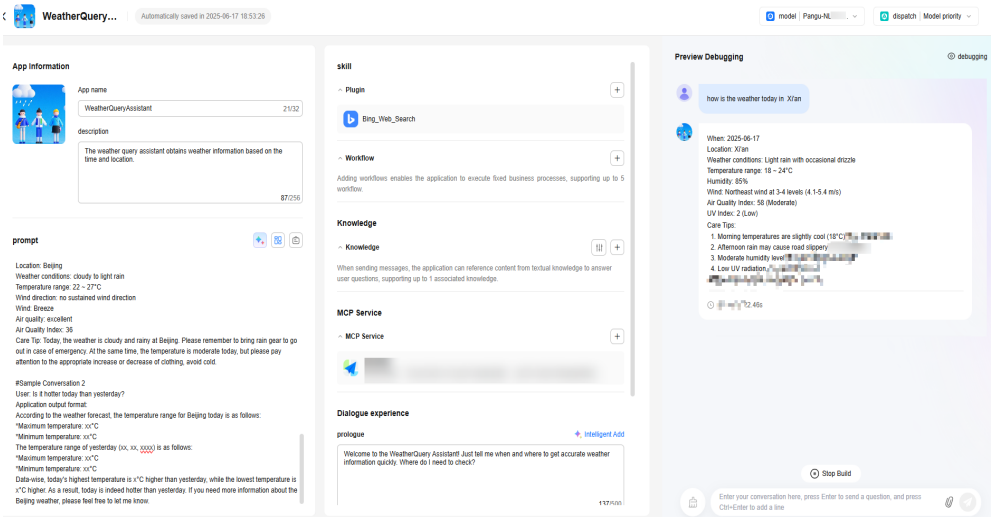
Step 3: In the **skill > MCP Service** area on the agent development page, click **Add** . The MCP service selection page is displayed.



Step 4: Select the required MCP service from the pre-installed services and personal services.

Step 5: Click **OK**.

Step 6: In the **Preview Debugging** area on the right, enter a question, for example, MCP, in the dialog box. View the call details.



- **Configure the MCP service during workflow application development.**

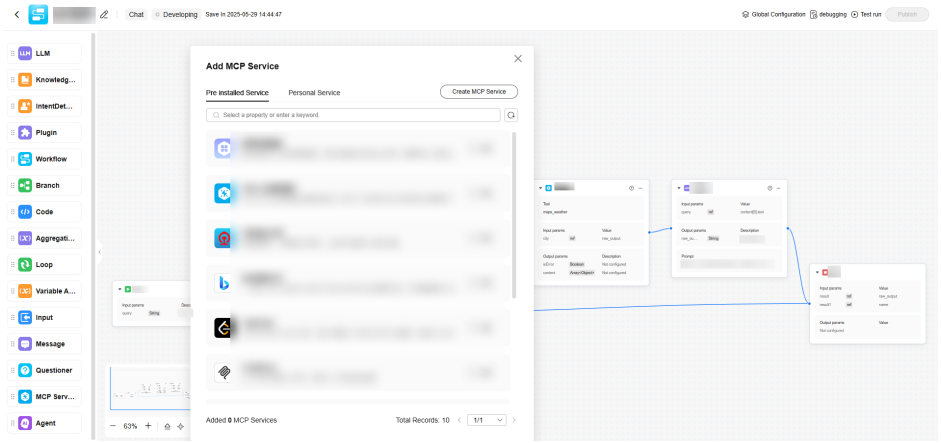
  Step 1: Log in to the agent development platform and choose **Workstation** from the navigation pane.

  Step 2: Click the **Workflow** tab page. Click **Create workflow**.

  Step 3: On the workflow development page, drag the MCP Service node in the navigation pane to the workflow orchestration canvas.

  Step 4: In the **Add MCP Service** dialog box, select the required MCP service from the pre-installed MCP services or personal services.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.



Step 5: Click **Add**.

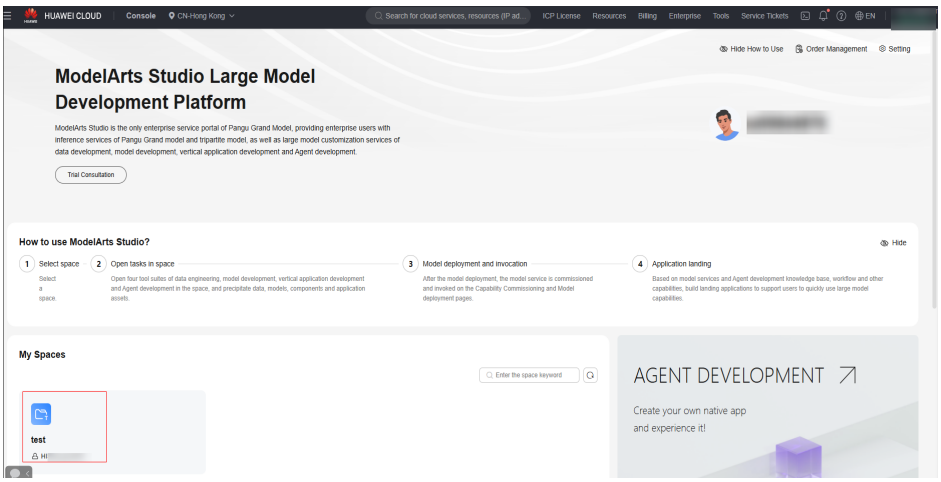
7.7.3 Managing MCP Services

You can edit and delete user-defined MCP services on the agent development platform.

Editing an MCP Service

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

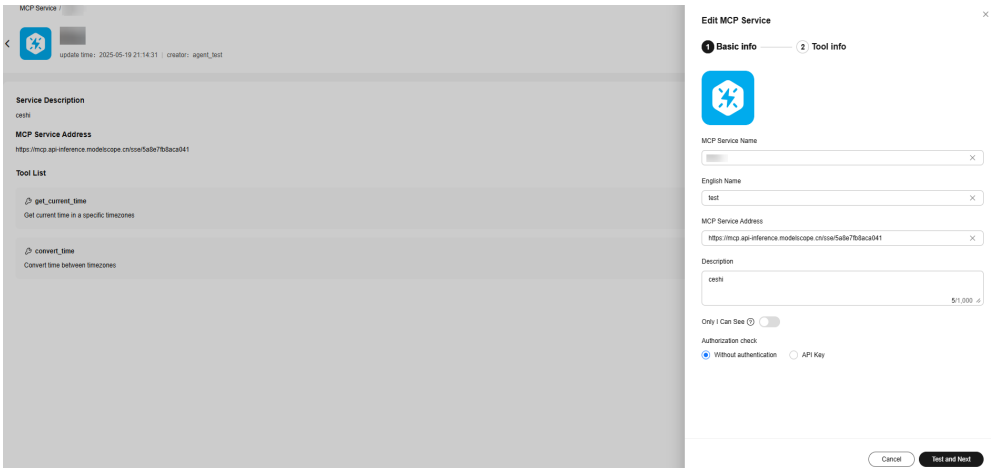
Figure 7-87 My Spaces



2. In the navigation pane, choose **Agent Dev**.
3. Choose **Workbench** > **MCP Service**.
4. Click the MCP service name. On the displayed details page, click **Edit** in the upper right corner.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

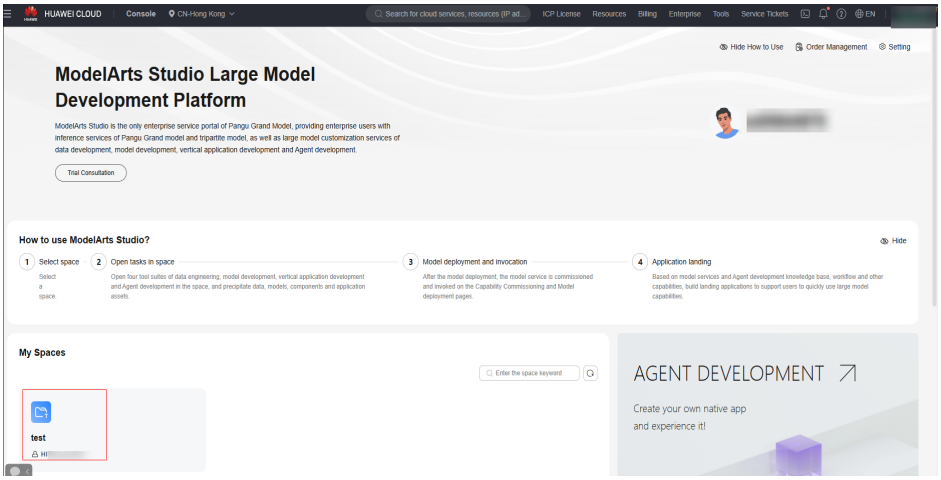
Figure 7-88 Editing an MCP service



Deleting the MCP Service

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

Figure 7-89 My Spaces



2. In the navigation pane, choose **Agent Dev**.
3. Choose **Workbench** > **MCP Service**.

- Click ··· and choose **Delete** to delete the current MCP service.

  NOTE

  Deletion is a high-risk operation. Before deleting an MCP service, ensure that the MCP service is no longer used.

7.7.4 Subscribing to the MCP Service

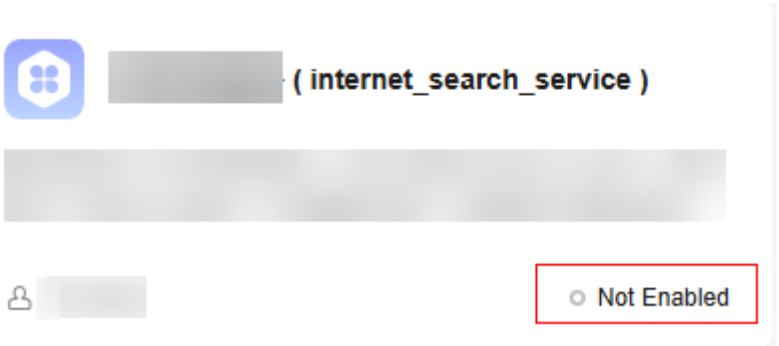
Some officially preset MCP services can be used only after being subscribed to.

1. Log in to the agent development platform.
2. Choose **MCP Market**.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

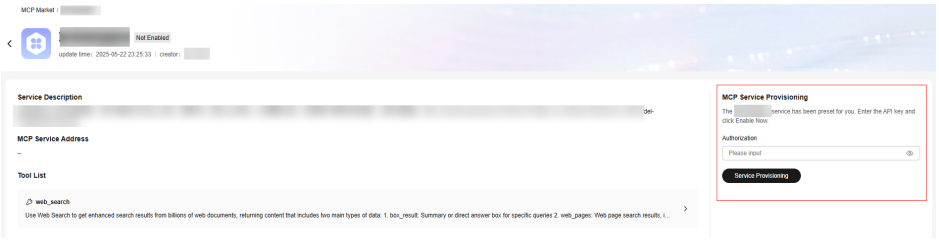
3. Check the service provisioning status, as shown in **Figure 7-90**.

Figure 7-90 Preset service provisioning status



4. Click the MCP service card. On the displayed page, enable the service, as shown in **Figure 7-91**.

Figure 7-91 Enabling the service



7.8 Common Errors and Solutions During Agent Development

Common Workflow Error Codes and Solutions

Table 7-30 lists common errors that occur at each node of the workflow, along with their corresponding solutions.

Table 7-30 Common errors and solutions at each node of the workflow

Module Name	Error Code	Description	Solution
Start node	101501	No value is transferred in the global configuration of the start node.	An error occurred on the start node. Contact customer service.
End node	101531	Failed to initialize the end node.	Check the configuration on the end node. The possible cause is that the verification failed.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Module Name	Error Code	Description	Solution
	101532	Template concatenation failed on the end node.	Check whether the template placeholder matches the input. Contact customer service.
	101533	Streaming processing on the end node failed.	Contact customer service.
LLM node	101561	Failed to initialize the LLM node.	Check the configuration of the LLM node. The possible cause is that the verification failed.
Code node	101591	Failed to initialize the code component.	Check the configuration on the code node. The possible cause is that the verification failed.
	101592	Failed to request the security sandbox on the code node.	Contact customer service.
	101593	Failed to execute the security sandbox on the code node.	Check whether the code syntax is correct and whether unreferenced variables are used.
	101594	Other errors are reported on the security sandbox of the code component.	Contact customer service.
	101595	Unknown error occurred during the execution on the code node.	Contact customer service.
Message node	101651	Failed to initialize the message component.	Check the configuration on the message node. The possible cause is that the verification failed.
	101652	The template information is missing on the message node.	Configure the prompt template of the message node.
	101653	The message node template is incorrectly concatenated.	Check whether the template placeholder matches the input. If the error persists, contact customer service.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Module Name	Error Code	Description	Solution
	101654	The message component failed to be executed.	Contact customer service.
	101655	The message component fails to be executed asynchronously.	Contact customer service.
Intent recognition node	101098	Failed to request the intent recognition prompt template.	Check whether the template placeholder matches the input.
	101097	The prompt for the intent recognition component to call the LLM does not comply with the model input specifications.	Check the format of the prompt, and the role and content of the message.
	101096	The intent recognition component failed to call the LLM.	Check whether the format and content of the message and the LLM service are normal.
	101095	The intent recognition component failed to parse the input or reference in the user query.	Check the format and content of the user query.
	101094	Failed to create the intent recognition prompt template.	Check the built-in template and the format and content of the system prompt.
Questioner node	101043	The number of dialog rounds in a single questioner exceeds the preset upper limit. The dialog status returns to the start node status.	You can increase the maximum number of dialog rounds.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Module Name	Error Code	Description	Solution
	101047	Failed to initialize the in-depth customization pre-processing and post-processing module.	Check whether the guardrail configuration meets the requirements.
	101048	The in-depth customization user reply rewriting (preprocessing) failed.	You can check the preprocessing guardrail code.
	101049	Failed to rewrite (post-process) parameter values generated by the in-depth customization model.	Check the post-processing guardrail code.
	101050	The default guardrail (time parameter parsing) fails to be executed.	Check whether the time type that can be processed exceeds the supported range.
	102053	The prompt template is incorrect.	Check whether the format of the prompt template is correct.
	103004	Model inference failed.	Check whether the model service is running properly.
Plug-ins	101741	Failed to initialize the plug-in engine.	Check the plug-in component configuration. The possible cause is that an error is reported during verification.
	101742	An error occurred when converting the parameter type of the workflow plug-in node.	Determine the name of the parameter that fails to be converted based on the error message and check whether the parameter type is correct.
	101743	The input of the workflow plug-in node does not exist in the plug-in definition.	Check whether the plug-in definition matches the corresponding component definition.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Module Name	Error Code	Description	Solution
	101744	The plug-in defines a response, but the actual plug-in execution result is inconsistent with the defined response.	Check whether the plug-in response definition matches the actual plug-in execution result.
	101745	An error occurs when the workflow plug-in node is executed.	If an error occurs during plug-in execution, locate the fault based on the error message. If the message does not contain valid information, the error is not captured.
	105001	An exception that cannot be captured occurs during plug-in execution.	Check whether the service is available.
	105004	Check param error during plug-in definition.	Determine the error parameter definition based on the error message.
	105005	The plug-in definition is invalid.	The data defined by the plug-in is invalid. For example, the field length exceeds the maximum. Check the error message.
	105008	The plug-in content is incorrect.	Contact customer service.
	105010	An authentication error occurs when the plug-in is running.	You can locate the authentication problem based on the error message and check whether the authentication information transfer and plug-in authentication definition are correct.
	105011	The response code returned during plug-in running is not 200.	You can view the actual HTTP return code based on the reported information.
	105012	The plug-in request times out.	The plug-in request times out. Check the plug-in service.
	105013	The result returned by the plug-in is too large.	Currently, 10 MB is supported. If the size exceeds 10 MB, an error is reported.
	105014	Plug-in request proxy error.	Check whether the plug-in service is faulty.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Module Name	Error Code	Description	Solution
Performing Authentication	110000	Authentication failed.	Check the authentication configuration.
	110001	Get users in domain in IAM failed	Check whether the user information is correctly configured.
Workflow	112501	Workflow authentication failed.	Check the authentication configuration.
	112502	Some required parameters are missing.	You can view the missing parameters in the printed logs.
	112503	The workflow fails to connect to the database.	Contact customer service.
	112504	Necessary permissions are missing.	Viewing the profile of the current user
	112513	An infinite loop occurs in the workflow.	Check the workflow canvas.
	112514	The workflow is referenced and cannot be deleted.	Check whether the workflow is referenced by the knowledge-based application.
	112600	Failed to convert the workflow IR.	Check whether the workflow configuration is correct.
	112941	Failed to obtain the workflow dialog history.	Contact customer service.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

8 Managing Workspace Assets

8.1 Introduction to Pangu Model Workspace Assets

Workspace assets in ModelArts Studio include data assets and model assets. These assets provide a basis for centralized management and efficient operations, facilitating unified query and operation management.

- **Data assets**: Published datasets are stored in the workspace assets as data assets. You can view the details about datasets, including the data format, size, and ratio. In addition, the platform supports management operations such as dataset deletion, enabling you to manage dataset resources in a unified manner. In this way, dataset resources can be flexibly invoked during model training and analysis, ensuring the standardization and security of data assets.
- **Model assets**: The model assets provided by the platform include preset models and published trained models. All these models are stored in space assets for unified management. You can view the historical versions and operation records of a preset model and perform further operations on the model, including training, compression, and deployment. In addition, the platform supports the export and import of Pangu models so that users can migrate Pangu models from other sites to the local site to facilitate model resource sharing.

8.2 Managing Pangu Data Assets

Introduction to Data Assets

Data assets refer to datasets that are managed, stored, and available on the platform.

Data assets are classified into the following types:

- **User-published datasets**

  You can publish a dataset as a data asset by choosing **Data Engineering** > **Data Publishing** > **Publish Task**. You can view details about published datasets, edit and delete datasets, and publish datasets to AI Gallery.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

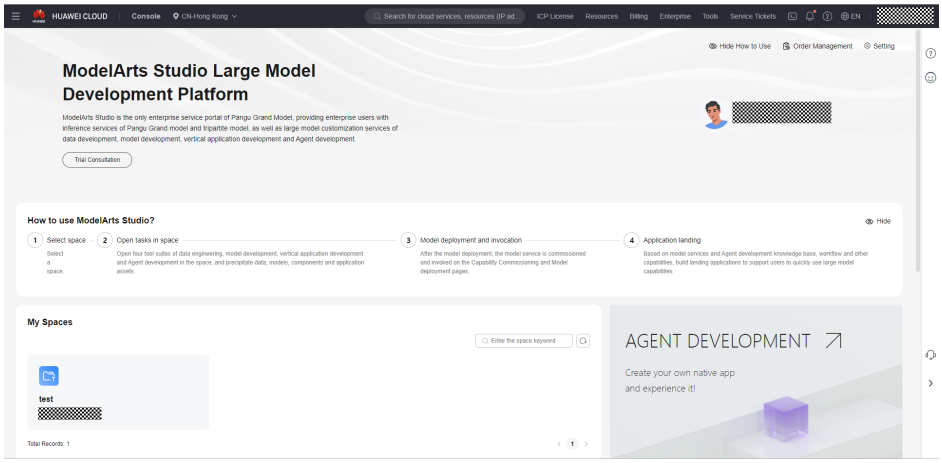
- **Data assets subscribed from AI Gallery**

  In addition to datasets published by users, the platform also allows you to subscribe to data assets from AI Gallery. AI Gallery enables the sharing of AI assets such as models, datasets, and AI applications. In this community, enterprises, individual developers, and other groups can securely share and purchase AI assets.

Publishing Data Assets to AI Gallery

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 8-1 My Spaces

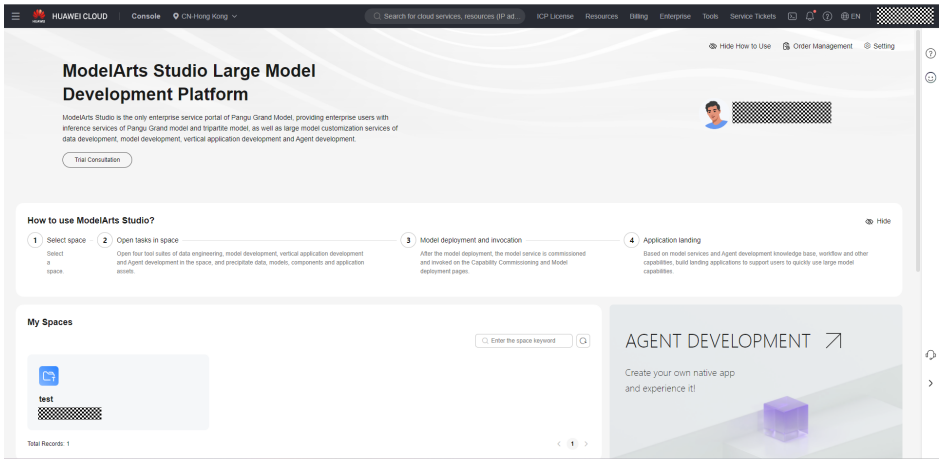
2. In the navigation pane, choose **Workspace Assets** > **Data**.
3. Click **Publish to Gallery** in the **Operation** column of the data asset in **Unpublished** state to publish the data asset.
4. In the **Publish to AI Gallery** dialog box, enter the AI Gallery asset name and description, available subscription area constraint, visible scope, and click **OK** to publish the data asset to AI Gallery.
5. The data asset publishing status is displayed on the data asset list page.
   - If the status is **Publishing**, the asset is being synchronized to AI Gallery.
   - If the status is **Published**, the asset has been synchronized to AI Gallery. You can click **View Publishing Information** in the **Operation** column to view the publishing information of the asset.
   - If the status is **Publish failed**, the asset fails to be synchronized to AI Gallery. You can click **Publish** to AI Gallery to publish the data asset again.

Subscribing to Data Assets from AI Gallery

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 8-2 My Spaces



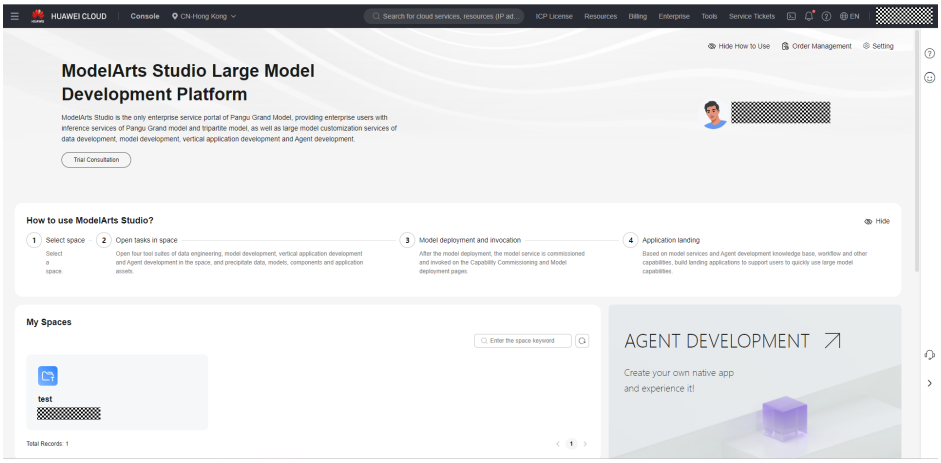
2. In the navigation pane, choose **Workspace Assets** > **Data**.
3. Click **Subscribe to Data** in the upper right corner. In the **Subscribe from AI Gallery** dialog box, select the data assets to be subscribed to and click **Next Step**.
4. Enter the asset name and description, and click **OK** to subscribe to the data asset.
5. The data asset publishing status is displayed on the data asset list page.
   - If the status is **Subscribing**, the asset is being synchronized from AI Gallery. Wait patiently.
   - If the status is **Subscription succeeded**, the asset has been successfully subscribed to from AI Gallery. You can click **View Subscription Information** in the **Operation** column to view the subscription information of the asset.
   - If the status is **Subscription failed**, the asset fails to be subscribed to from the AI Gallery. You can click Re-subscribe to subscribe to the data asset from the AI Gallery again.
6. The subscribed data assets are displayed on the **Data Engineering** > **Data Management** > **Datasets** > **Original Dataset** page. You can process and publish the data assets.

Managing Data Assets

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 8-3 My Spaces



2. In the navigation pane, choose **Workspace Assets** > **Data**.
3. Click the **Task Publishing** tab and perform the following operations on the datasets published by users:
   - View basic information. You can click a data asset to view its basic information such as ratio details and data details.
   - Publish data assets to AI Gallery. Click **Publish to AI Gallery** in the **Operation** column to publish the data asset to AI Gallery.
   - View the publishing information. Click **View Publish Information** to view the publishing information of the asset. (This operation can be performed only on the data assets that have been published to AI Gallery.)
   - Editing properties. Choose **More** > **Edit Property** in the **Operation** column to edit the name, description, and visibility of the data asset.
   - Delete a data asset. Choose **More** > **Delete** in the **Operation** column to delete the current data asset.
   - Unpublish data assets from AI Gallery. Choose **More** > **Cancel Publish** to AI Gallery in the **Operation** column to unpublish the data asset from AI Gallery. (This operation can be performed only on the data assets that have been published to AI Gallery.)
4. Click the **AI Gallery** tab and perform the following operations on the data assets subscribed from AI Gallery:
   - View the subscription information. Click **View Subscription Information** in the **Operation** column of a specific data asset to view the subscription information such as the asset name and description.
   - Edit properties. Choose **More** > **Edit Properties** in the **Operation** column to edit the name, description, and visibility of the data asset.
   - Delete a data asset. Choose **More** > **Delete** in the **Operation** column to delete the current data asset.
   - Re-subscribe to a data asset. If the subscription fails, click **Resubscribe** in the **Operation** column to subscribe to the required data asset from AI Gallery again.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

8.3 Managing Pangu Model Assets

Introduction to Model Assets

Models that can be used for trial, subscribed to, or published after training are regarded as model assets and stored in space assets, facilitating unified management and operations. You can view all historical versions and operation records of a model to track the model evolution. In addition, the platform supports a series of convenient operations, including model training, compression, and deployment, thereby simplifying the model development and application process. These functions help users efficiently manage the model lifecycle and improve asset management efficiency.

Model assets are classified into the following types:

- **Preset models**

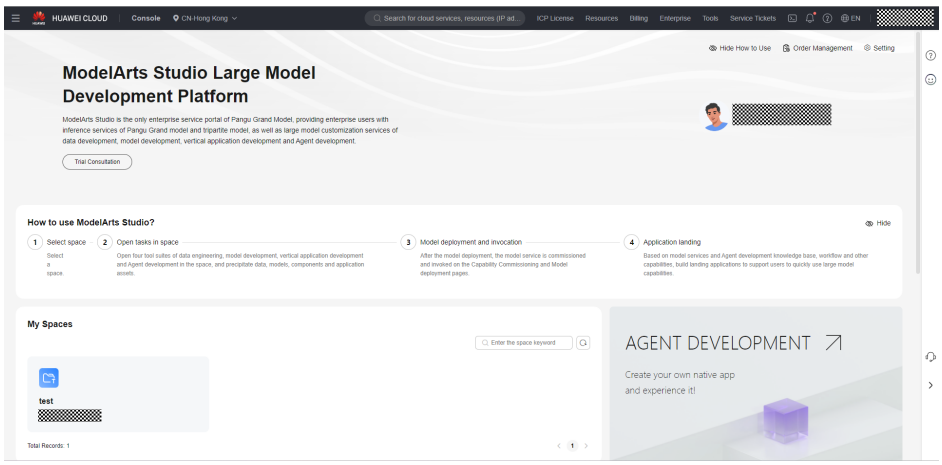
  Preset models that can be tried or have been subscribed to on the platform.
- **Models published by users**

  You can publish a trained model as a model asset. You can view details about a published model, edit attributes, and delete the model.

Managing Pangu Model Assets

1. Log in to ModelArts Studio Large Model Deveopment Platform. In the **My Spaces** area, click the required workspace.

   Figure 8-4 My Spaces

2. In the navigation pane, choose **Workspace Assets** > **Models**.
3. Click the **Other Spaces** tab. In the model list, click a model and perform the following operations on it:
   - View historical versions of the model. On the **Version List** tab page, you can view each version of the model.
   - Perform operations such as training, compressing, or deploying the model. On the **Version List** tab page, you can train, compress, or deploy models of different versions. Click the corresponding button to go to the related operation page.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

  - View operation records. On the **Operation Record** tab page, view the operation records of the current model.

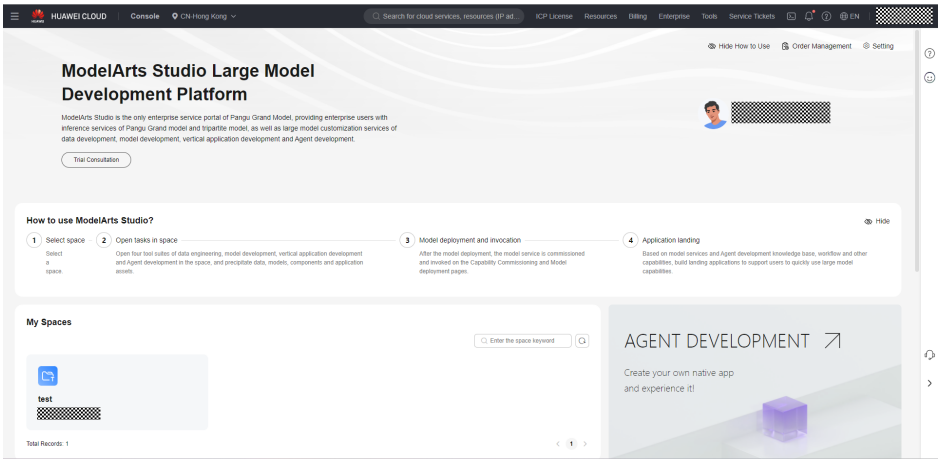
- Click the **Own Space** tab and perform the following operations on models published in the current workspace:
  - View model information. Click a model name. On the model information page that is displayed, view the basic information and operation records of the model.
  - Editing properties. Click **Edit Property** in the **Operation** column to modify the model asset name, description, and asset visibility.
  - Perform operations such as training, compressing, or deploying the model on the model list page. Click the corresponding button to go to the related operation page.

8.4 Managing Pangu Component Assets

Managing Component Assets

1. Log in to ModelArts Studio. In the **My Spaces** area, click the required workspace.

   Figure 8-5 My Spaces

2. In the navigation pane, choose **Workspace Assets** > **Components** to view component assets in the current workspace, other workspaces, and preset workspaces.
3. Click the **Own Space** tab to edit or delete the components published in the current workspace.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

9 Managing Resource Pools

9.1 Creating an Edge Resource Pool

Edge deployment refers to deploying models on edge devices. These devices are usually servers purchased by users. ModelArts manages these servers as edge resource pools, and models can be deployed in these edge resource pools using the PanguLM service.

ModelArts edge nodes are devices provided by ModelArts for deploying edge services. Before creating an edge resource pool, you must create a ModelArts edge node. After a node is created, download the certificate and edge agent firmware, copy the firmware to the node, and run the registration command to register the device.

For details about how to create an edge resource pool, see **Table 9-1**.

Table 9-1 Creating an edge resource pool

Step	Description
Preparations	Preparations for creating an edge resource pool
Step 1: Registering an Edge Computing Node	Procedure for registering edge resource pool nodes
Step 2: Setting Up an Edge Server Cluster	Procedure for setting up an edge server cluster
Step 3: Installing the Ascend Plug-in	Guidance for installing the Ascend plug-in
Step 4: Creating a Certificate	Procedure for creating a certificate required for creating a load balancer
Step 5: Creating a Load Balancer	Procedure for creating a load balancer

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

NOTE

- Currently, only some models on ModelArts Studio support edge deployment. For details, see "Model Capabilities and Specifications" in *Service Overview*.
- To use the edge deployment function, you need to enable the edge resource pool function in ModelArts. This function is a **whitelist feature**. To enable this function, contact ModelArts customer service.
- **Creating an edge resource pool is complex. You are advised to contact Pangu customer service for assistance.**

Preparations

This section uses the **largemodel** cluster as an example to describe how to deploy models on edge services. The following table lists the example cluster information.

Table 9-2 Example cluster information

Cluster Name	Node Type	Node Name	Specifications	Remarks
largemodel	Controller	ecs-edge-XXXX	Kunpeng General computing | 8 vCPUs | 29 GiB | rc3.2xlarge.4 image EulerOS 2.9 64-bit with Arm for Tenant 20230728 base 2.9.15	Public IP address: 100.85.220.207 root password: / CPU architecture: AArch64 (Log in to the device and run the **arch** command to view the CPU architecture.)
	Worker	bms-panguXXXX	CPU:Kunpeng Memory: 24 x 64 GB DDR4 RAM (GB) Local disk: 3 x 7.68 TB NVMe SSDs Extended configuration: 2 x 100GE + 8 x 200GE Type: physical.kat2e.48xlarge.8.313t.ei.pod101 euler2.10_arm_sdi3_1980b_hc_sdi5_b080_20230831v2	Public IP address: 100.85.216.151 root password: / CPU architecture: AArch64 (Log in to the device and run the **arch** command to view the CPU architecture.)

1. Download the dependency package.
   - Docker download: **https://download.docker.com/linux/static/stable**

   Select the corresponding CPU architecture and download Docker 19.0.3 or later.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- K3S download: **https://github.com/k3s-io/k3s/releases/tag/v1.21.12%2Bk3s1**

Download the binary file and air-gap image based on the CPU architecture.

2. Install the NPU driver and firmware.

   Run the **npu-smi info** command to check whether the driver has been installed. If NPU information is displayed, the driver has been installed.

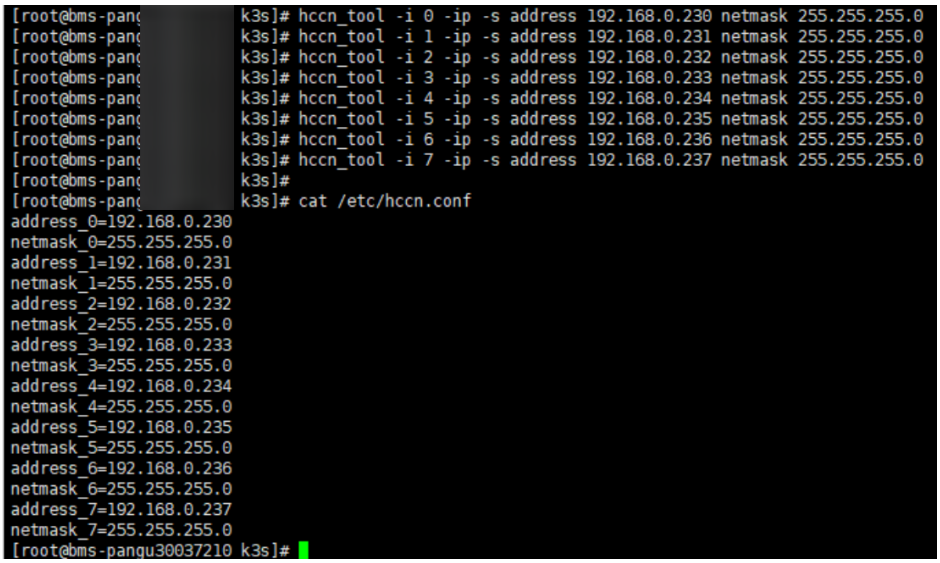
3. Run the **hccn_tool** command to view the NIC configuration.

   a. Run the following command and check whether NIC information is displayed. If yes, the NIC has been configured. Otherwise, go to the next step.
   ```
   cat /etc/hccn.conf
   ```

   b. Run the following command to check the number of NPUs:
   ```
   npu-smi info
   ```

   c. Run the following commands (replace the IP addresses with actual ones):
   ```
   hccn_tool -i 0 -ip -s address 192.168.0.230 netmask 255.255.255.0
   hccn_tool -i 1 -ip -s address 192.168.0.231 netmask 255.255.255.0
   hccn_tool -i 2 -ip -s address 192.168.0.232 netmask 255.255.255.0
   hccn_tool -i 3 -ip -s address 192.168.0.233 netmask 255.255.255.0
   hccn_tool -i 4 -ip -s address 192.168.0.234 netmask 255.255.255.0
   hccn_tool -i 5 -ip -s address 192.168.0.235 netmask 255.255.255.0
   hccn_tool -i 6 -ip -s address 192.168.0.236 netmask 255.255.255.0
   hccn_tool -i 7 -ip -s address 192.168.0.237 netmask 255.255.255.0
   ```

   d. Run the **cat /etc/hccn.conf** command. If the following NIC information is displayed, the configuration is complete.



4. Configure the NFS web disk service.

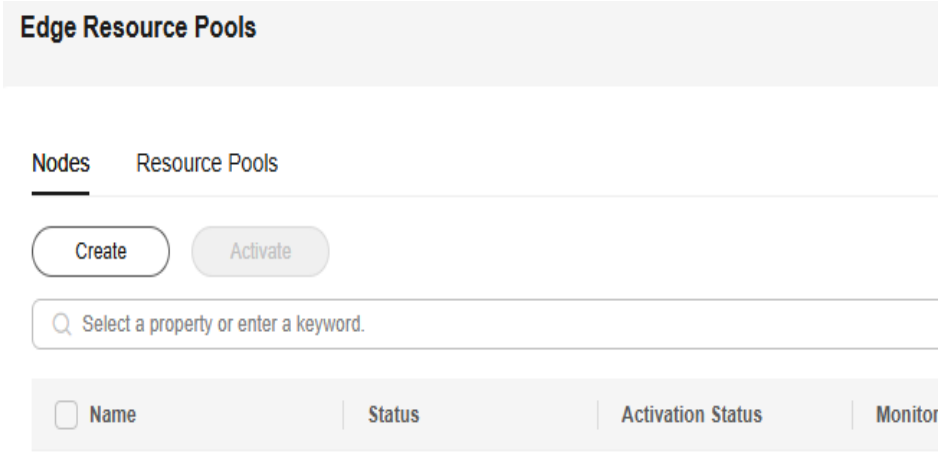
   If a large model is deployed as mirrored models in different nodes, one of them needs to provide Network File System (NFS), which allows you to access the model by mounting NFS.

Step 1: Registering an Edge Computing Node

1. Go to the ModelArts console and select the required workspace.
2. In the navigation pane, choose **Resource Management** > **Edge Node**. On the **Nodes** tab page, click **Create**.

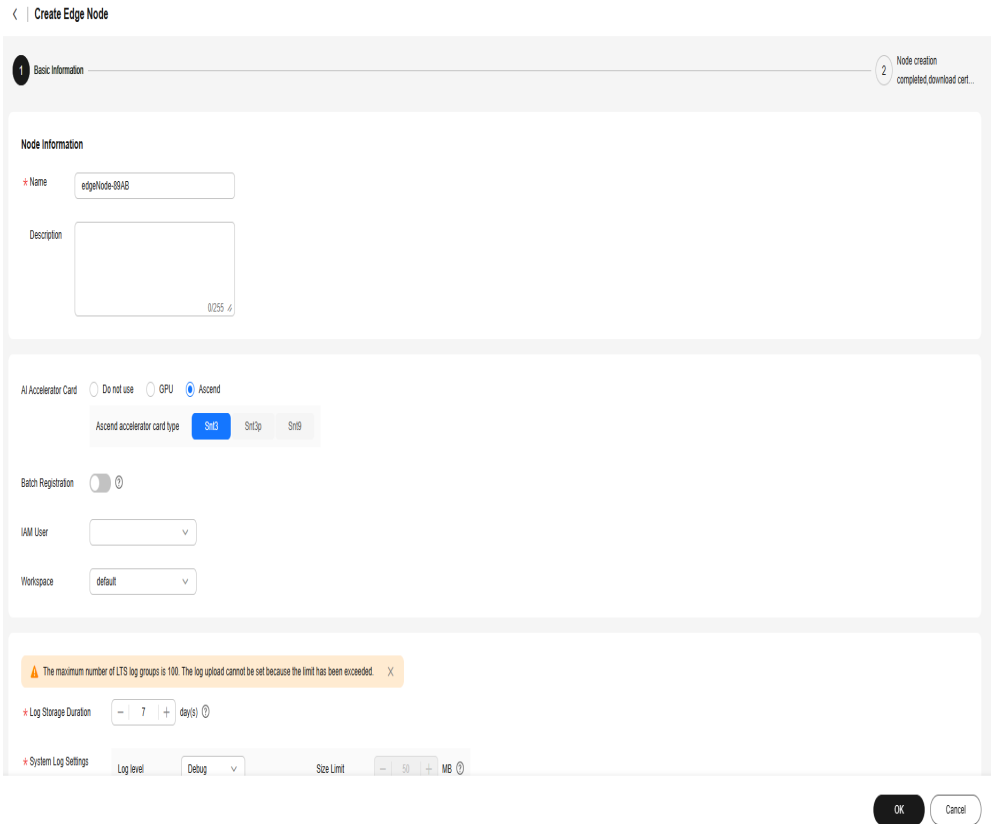
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 9-1 Edge Node



3. On the **Create Edge Node** page, enter the node name and configure the AI accelerator card and log information.
   - If the node has NPUs, select **Ascend** for **AI Accelerator Card** and select the accelerator card type.
   - If the node does not have an accelerator card, set **AI Accelerator Card** to **Do not use**.

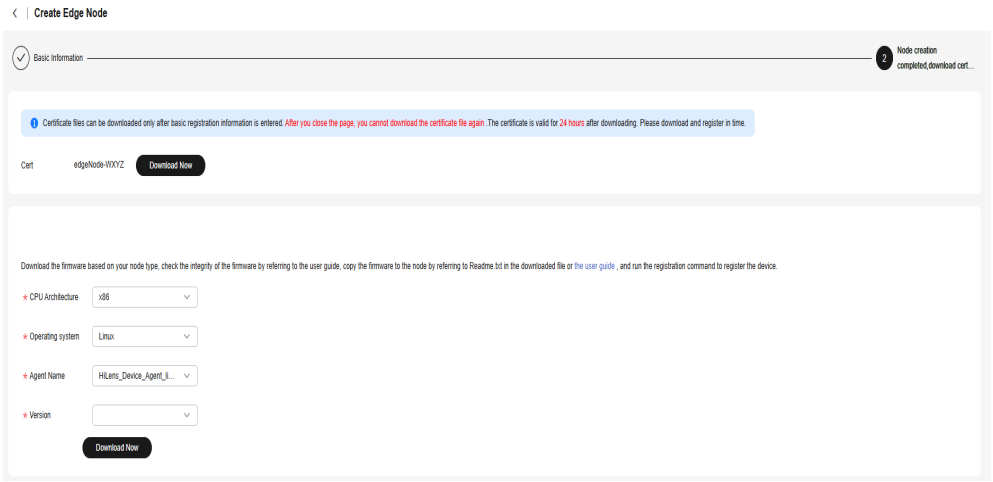
Figure 9-2 Configuring edge node parameters



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

4. After the basic information is configured, click **OK**. Click **Download Now** to download the device certificate and Agent firmware, and rename the device certificate and Agent firmware as **license.tgz** and **hilens-agent.tgz**, respectively.

Figure 9-3 Downloading the required certificate



Step 2: Setting Up an Edge Server Cluster

1. Run the following command to generate a Docker certificate:

```
bash cluster_install-ascend.sh generate_docker_cert --pkg-path=/home/hilens/pkgs
```

NOTE

Note that this command needs to be executed only once. If the related certificate exists, skip this step.

2. Save the downloaded files in the following directories based on **Preparations** and **Step 1: Registering an Edge Computing Node**. Modify the names of the downloaded files. The **certs** certificate in Docker is automatically generated and does not need to be modified.

```
pkgs // Package directory, which is user-defined.
    docker
        docker.tgz    // Docker binary file. The version must be later than 19.0.3.
        certs         // Certificate generated by running the generate command. After --pkg-path is specified, the certificate is automatically created in the certs directory.
            ca.crt
            server.crt
            server.key
    k3s
        k3s // Executable k3s file
        agent
            images
                k3s-airgap-images-[arm64|amd64].tar.gz // K3s offline image
    hilens-agent
        hilens-agent.tgz // HiLens Agent firmware package
        license.tgz // HiLens device license
```

3. Run the following command on the worker node:

```
bash -x cluster_install-ascend.sh --pkg-path=/home/hilens/pkgs --node-type=worker --host-ip=192.168.0.209
```

Run the following command on the master node:

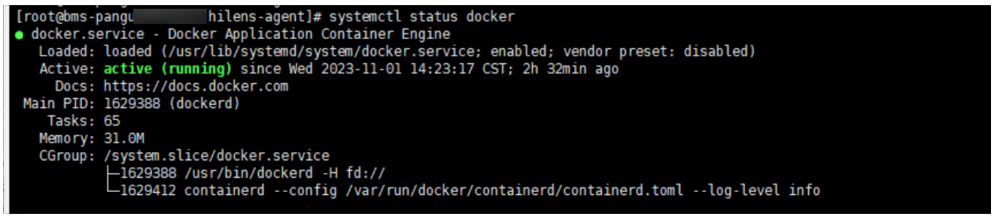
```
bash -x cluster_install-ascend.sh --pkg-path=/home/hilens/pkgs --node-type=controller --host-ip=192.168.0.150
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- **cluster_install-ascend.sh** is used to install Docker, HDAD, and K3S. Contact Pangu customer service to obtain it.
- **pkg-path** is the directory of the installation package file integrated in step 2.
- **host-ip** indicates the IP address of the device in the cluster. Generally, it is an internal IP address.
- **node-type** indicates the cluster node type. **worker** indicates a worker node, and **controller** indicates a master node.

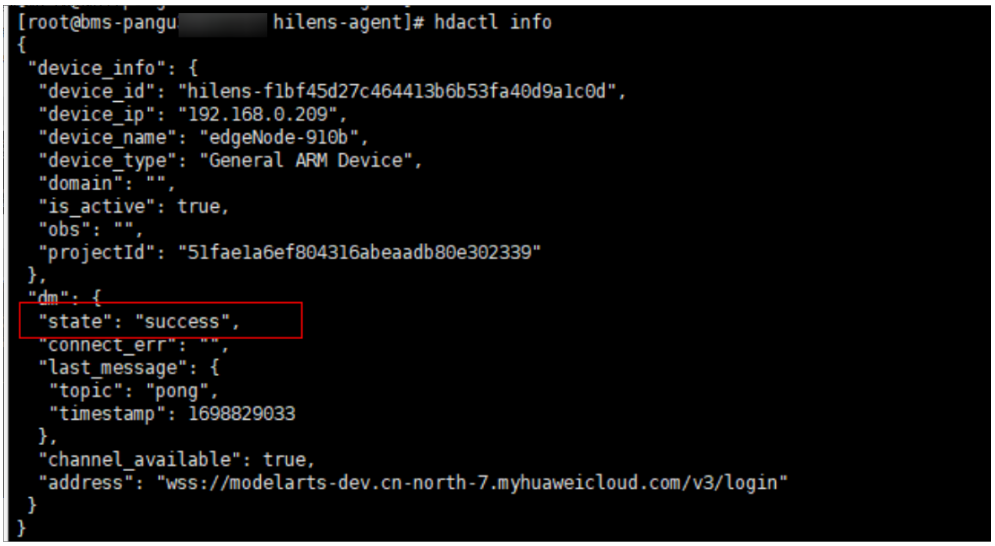
4. Run the following command on the server to check whether Docker is installed:

```
systemctl status docker
```



5. Run the following command on the server to check whether the edge agent is installed:

```
hdactl info
```



6. Configure the NFS web disk service.
   a. **Install NFS.**

      Connect to the Internet and download the dependencies.

      - **Ubuntu**

        Online installation:

        ```
        sudo apt install nfs-kernel-server
        ```

      - **EulerOS**

        Online installation:

        ```
        sudo yum install nfs-utils
        ```

      NOTE

      If you need to install it offline, contact Pangu customer service.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

- (Optional) Enable the rpc-bind and nfs services on the firewall:
  ```
  sudo firewall-cmd --zone=public --permanent --add-service={rpc-bind,mountd,nfs}
  sudo firewall-cmd --reload
  ```

b. **Create a shared directory on the web disk.**

The storage space of this path should be sufficient for storing large model files. Set **/var/docker/hilens** as the root directory of the web disk to access it in the container. Run the following command:

```
sudo mkdir -p /var/docker/hilens
```

Run the following command to set the access permission of the path to **1000:100** (consistent with the configuration of **/etc/exports**):

```
sudo chmod 750 /var/docker/hilenschown -R 1000:100 /var/docker/hilens
```

Run the following command to view the permission:

```
ls -l /var/docker | grep hilens
```

c. **Add the web disk access permission.**

Run the following command to configure the nfs-server access whitelist and the path of the web disk shared file:

```
sudo vim /etc/exports
```

Add the following configurations:

```
/var/docker/hilens 172.xxx.0.0/24(rw,no_all_squash,anonuid=1000,anongid=100,fsid=0)
```

172. xxx.0.0/24 is the internal IP network segment of the cluster. To view the IP address, log in to the master node and run the **hdactl info** command. For example, if the queried IP address is **172.16.0.22**, you can set the IP address to **172.16.0.0/24**.

In this configuration:

- **/var/docker/hilens**: root directory of the web disk.
- **192.168.0.0/24**: client IP address range. All nodes whose IP addresses are in this range can access **/var/docker/hilens**. The asterisk (*) indicates that there is no restriction. You can also enter the IP address of a specific node.
- **rw** indicates read and write permission.
- **anonuid** indicates the ID of the mapped anonymous user, and **anongid** indicates the mapped anonymous user group, that is, the file owner displayed in the container after the file is mounted to the container.
- **no_all_squash**: Common users can be authorized to use the files.

Run the **:wq** command to save the settings and update the NFS configuration. Then, run the following command:

```
exportfs -rv
```

d. **Start NFS and rpcbind.**

Set the service to start upon system startup.

```
systemctl enable nfs-server && systemctl enable rpcbind && systemctl start rpcbind nfs-server
```

Run the following command to check whether the preceding configuration is correct. If the following information is displayed, the configuration is correct, that is, the NFS service is installed.

```
showmount -e localhost
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

e. **(Optional) Verify the NFS configuration.**

Create a directory on the non-NFS service node.

```
sudo mkdir ~/data
```

Mount an NFS server:

```
sudo mount -t nfs 192.168.xx.xxx:/var/docker/hilens ~/data
```

After the mounting, run the following command to view the result:

```
mount
```

The mounting is successful if the following information is displayed:

```
...
...
192.168.0.150:/var/docker/hilens on ~/data type nfs4
```

f. **Test the NFS function.**

On the client, create a file in the shared directory.

```
cd ~/data
sudo touch a
```

View the created file on the NFS server whose IP address is 192.168.0.150.

```
cd /var/docker/hilens
ls -l
```

g. **Configure a Yum repository.**

If **yum install** is running properly, skip this section.

i. Configure the Yum repository address.

- Backing up Yum configurations

```
mkdir -p /etc/yum.repos.d/bak/ mv -f /etc/yum.repos.d/*.repo /etc/yum.repos.d/bak/
```

- x86 EulerOS configuration

```
cat> /etc/yum.repos.d/his-mirrors.repo<<"EOF" [EulerOS_2.10_base] name=EulerOS_2.10_base baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/os/x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1 [EulerOS_2.10_devel_tool] name=EulerOS_2.10_devel_tool baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/devel_tools/x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1 [EulerOS_2.10_updates] name=EulerOS_2.10_updates baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/updates/x86_64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=0 enabled=1 EOF yum clean all && yum makecache > /dev/null 2>&1
```

- Arm EulerOS configuration

```
cat> /etc/yum.repos.d/his-mirrors.repo<<"EOF" [EulerOS_2.10_base] name=EulerOS_2.10_base baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/os/aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1 [EulerOS_2.10_devel_tool] name=EulerOS_2.10_devel_tool baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/devel_tools/aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1 [EulerOS_2.10_updates] name=EulerOS_2.10_updates baseurl=http://his-mirrors.huawei.com/install/euleros/2.10/updates/aarch64/ gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS gpgcheck=1 enabled=1 EOF yum clean all && yum makecache > /dev/null 2>&1
```

ii. Configure a DNS server.

```
cat >> /etc/resolv.conf<<"EOF" nameserver 10.189.32.59 nameserver 10.72.55.103 nameserver 10.98.48.39 EOF
```

Certificates are automatically downloaded. However, if the certificates are not automatically installed on some servers or or if there are issues with the installed certificates, you can manually install them by running the following command.

```
 cat > /etc/pki/rpm-gpg/RPM-GPG-KEY-EulerOS <<"EOF" -----BEGIN PGP PUBLIC KEY BLOCK-----  mQENBFx4vK4BCACrq4PA8EZEr5XH08bfh9rlms +QDZsJYhqqlXWx4qsZ8dAqdWLH
```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

```
O2Dm0HYk17xVwTZjXUYH9rz1gn2bGa5At4xTpH7FHMDpNG8DfwC6UpMKEmGGvy/S
OfL4fl6Yq2tCHAx3LrHXO9PGigafz5XMDtBySI4ixOR4M/w8IaIPEyN0BfGC6DQ5
PZlXMPDOc5VW1NhxwH0u0yHH1lLITKBAEifTQa8+3YZY54PWvBbYxOcCCS63FOTn
pk0wWuwm2JqozDBxV/w8Ty0c25+y4FTiUGzOj2e/3K1ls+Zs/tTxf8asKpFH/dYN
ffXmdkPLDCHRGkTyrLPdghoFmVg4XhrOhsf/ABEBAAG0LkV1bGVyT1MgKEV1bGVy
T1MgMi4wIFNQOCkgPGV1bGVyb3NAaHVhd2VpLmNvbT6JAVQEEwEIAD4WIQS4wSv5
6CWqOEUVIHBxp+kP4ux1vAUCXHi8rgIbAwUJA8JnAAULCQgHAgYVCgkICwIEFgID
AQIeAQIXgAAKCRBxp+kP4ux1vN/UB/4jy2kRiFznSXfWz5SX0szLOf1FgDssdRZG
xASHonAJqrV19mkG2pNTkgip9LZQsCqLbxj5FV+TMm1o+6jubd9qRPePlV2Tpc0T
m3cDmpcZbW/XrFh4dLdN644TwKDAPcK8TK/wOepFVtjhx1Qc8o+8nuoFhmsMoKkf
AA1DYDDCZbpbgMQwMV3yKh002CFRlCnVMylqOi8U+FrYfphnsYfujXpKu9g+FmO6
ju0xVhxyFOVCEicamKiel3ZS9z06+PIFL3KP/nKC7pu0tfaxogJjCh8y+ZIF/FJU
ygHpZYKQJyJNO8Do7AucudbruOxqGqhD2BIBtX1JNP/hkKj4w+OJuQENBFx4vK4B
CAD1SYnEWn0mf3umbucovVYHaywJqhErB3ia+Ykq3InKvIORf1reiCRvVuse2wZr
cnPWEKeRE7tTEIZ5hCuLYQzaqngqVkwqLbRR6vtxiDhTWNgJH9+GuokgjotQ3/7T
AXH7AwG57OPp6vvaiazCDjhy5t3Vr1snWkiwWkJR2GFRkuwKu7FDLjc1n2dx4zLF
zRZJa5TTAR1zrHWgVwkLxgq0+eJHWq7eHFw1SBjmc4Vs4z/QI+Q+3rkVBiqCmr11 /
XQz9ZePdOl/8fCuNh4l480c7AFlFt8IvKBP6Kh+jShxbED2NJpRV04MhV6SyA69 ixd/
VKtPJMRcRg3phyuuCedrABEBAAGJATwEGAEIACYWIQS4wSv56CWqOEUVIHBx p
+kP4ux1vAUCXHi8rgIbDAUJA8JnAAAKCRBxp+kP4ux1vBGeB/4ubYvxZ6/apb+i
MCtRIuA15PWEwVFTVfKirvEliY4fAjw5HSlfrnN4FV/OCTIRHecuNBNBfLt78DoK
08x7fYtEBqIN6pDanjsSwbPhuzhz6m4C/GWLqqDi8SCaVTQsJqKc2QHjr7CaBIuo GRlB84/pOq
+kGAnMZPhCjy52K9x5zRpq6zTUpV5XPeLCBn6Kc8GW1Lk6K1eXsn09
KdIhb9JqTDtQx0eOS1p0fJlTb68Pj406IYJ16FaXmMTcYvep6HhVATQGbPDuIepd
BK12nQEDrezkGR5vH9nMraQuZTvADuRFFgQvZQ5QMYAzMa0RhQTHKMOwCMq8mBq7
3Csgrxwa =kXkt -----END PGP PUBLIC KEY BLOCK----- EOF
```

7. Configure the **hda.conf** configuration file.

   a. Log in to the microservice node and run the following command:
   ```
   vi /etc/hilens/hda.conf
   ```

   b. Add the following configurations:
   ```
   hilens.nfs.server.ip=192.168.0.150
   hilens.nfs.mount.dir=/home/mind/model
   hilens.nfs.source.dir=/var/docker/hilens
   ```

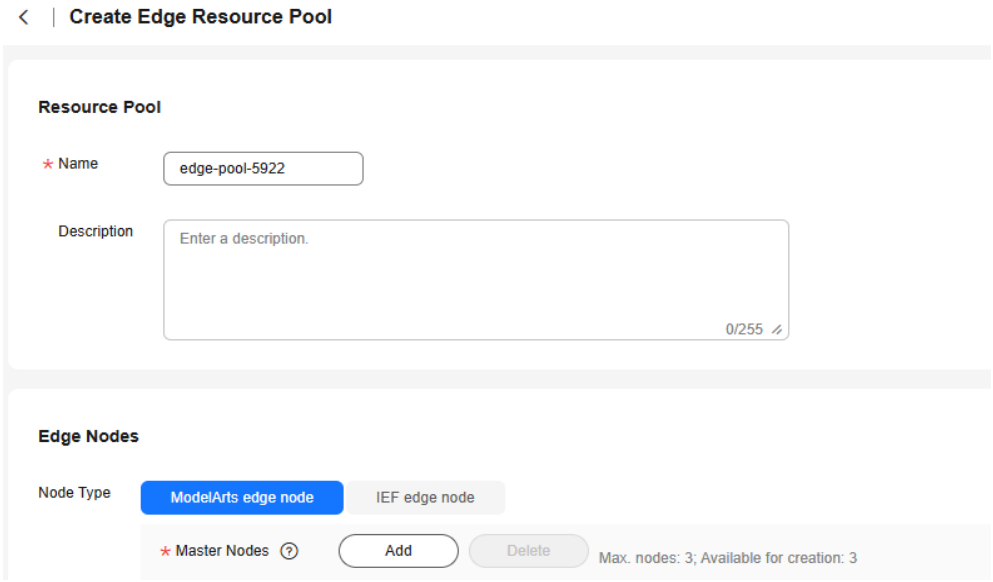
   In the preceding command, **server.ip** indicates the internal IP address of the NFS storage node, **mount.dir** indicates the default mount directory of the large model, and **source.dir** indicates the download directory of the large model.

   c. After the configuration is complete, run the following command to restart the firmware:
   ```
   systemctl restart hdad
   ```

8. Go to the ModelArts console and select the required workspace. Choose **Edge Resource Pools** > **Nodes** and click **Activate** in the **Operation** column of the current node. Check that the node status changes from **Inactive** to **Active**.
9. Choose **Edge Resource Pools** > **Resource Pools** and click **Create**. Enter the resource pool name, select **ModelArts Edge Node**, click **Add** next to **Master Node**, select the master node to be added, and click **OK**.

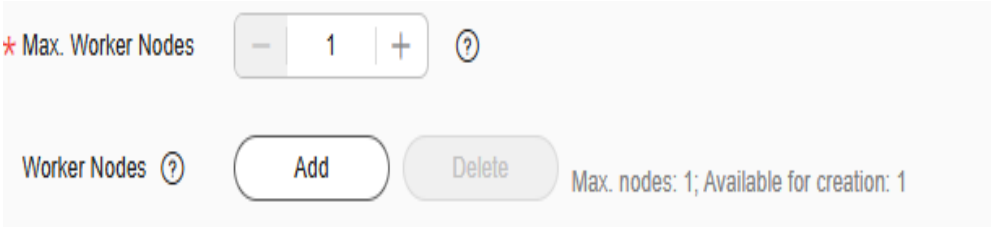
 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 9-4 Adding a master node

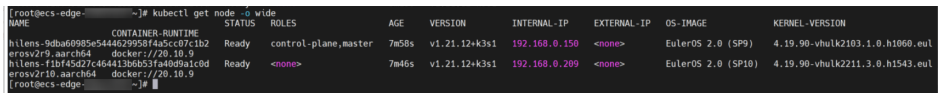


10. In the **Worker Nodes** area, click **Add**, select the worker node to be added, and click **OK**.

Figure 9-5 Adding a Worker node



11. Click **Create Now**. You can view the node status in the resource pool list. If the status is **Running**, the creation is successful.
12. Run the following Kubernetes command on the master node to verify the edge pool creation:
    a. Run the following command to set up a connection:
    ```
    ln -s /home/k3s/k3s /usr/bin/kubectl
    ```
    b. Run the following command to query the node status:
    ```
    kubectl get node -o wide
    ```
    c. If the status of all nodes is **Ready**, the cluster is created.



Step 3: Installing the Ascend Plug-in

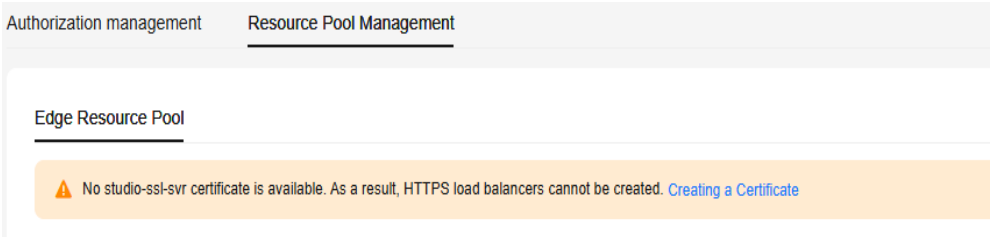
For details, visit **https://www.hiascend.com/document/detail/en/mindx-dl/300/productdescription/dlug_description_01_000001.html**.

Step 4: Creating a Certificate

If no certificate is available on the **Edge Resource Pools** tab page, as shown in **Figure 9-6**, create one by referring to the following steps.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

Figure 9-6 No certificate available

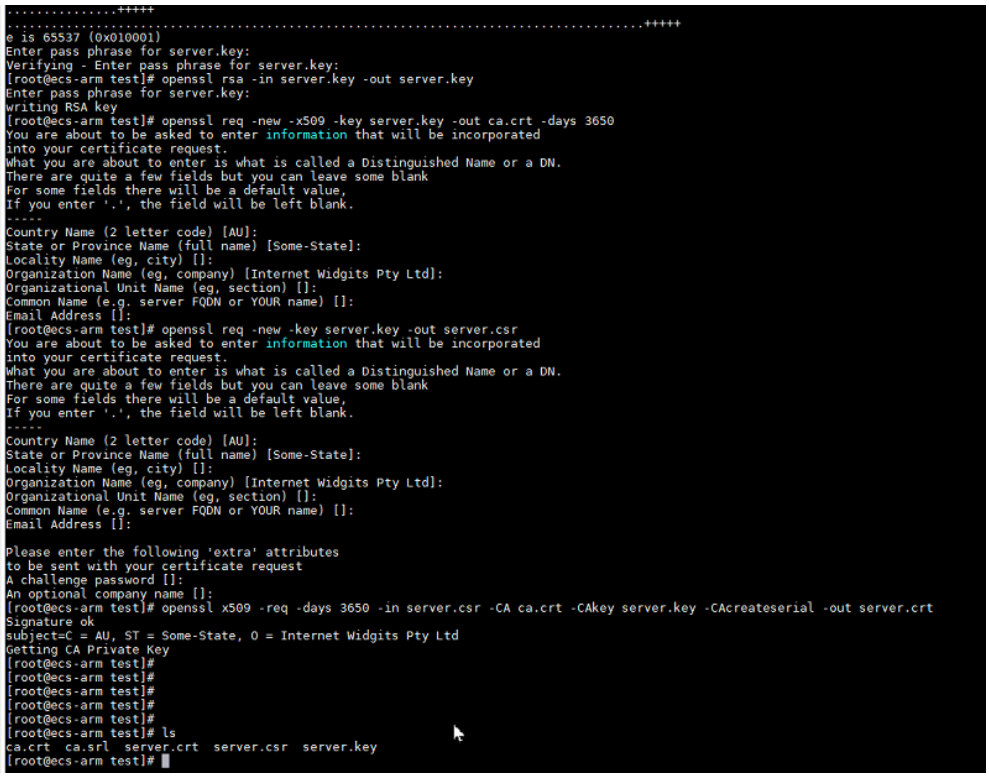


1. Prepare a Linux server where OpenSSL has been installed and run the following commands to create a certificate:

   When you run the command, the system prompts you to enter a password that contains at least four characters, for example, 123456. Remember the password, which will be used in subsequent steps.

   ```
   1. Generate the server.key file.
   openssl genrsa -des3 -out server.key 2048
   2. Generate an unencrypted server.key file.
   openssl rsa -in server.key -out server.key
   3. Generate the ca.crt file.
   openssl req -new -x509 -key server.key -out ca.crt -days 3650
   4. Generate the server.csr file.
   openssl req -new -key server.key -out server.csr
   5. Generate the server.crt file.
   openssl x509 -req -days 3650 -in server.csr -CA ca.crt -CAkey server.key -CAcreateserial -out server.crt
   ```

   Figure 9-7 Command execution example

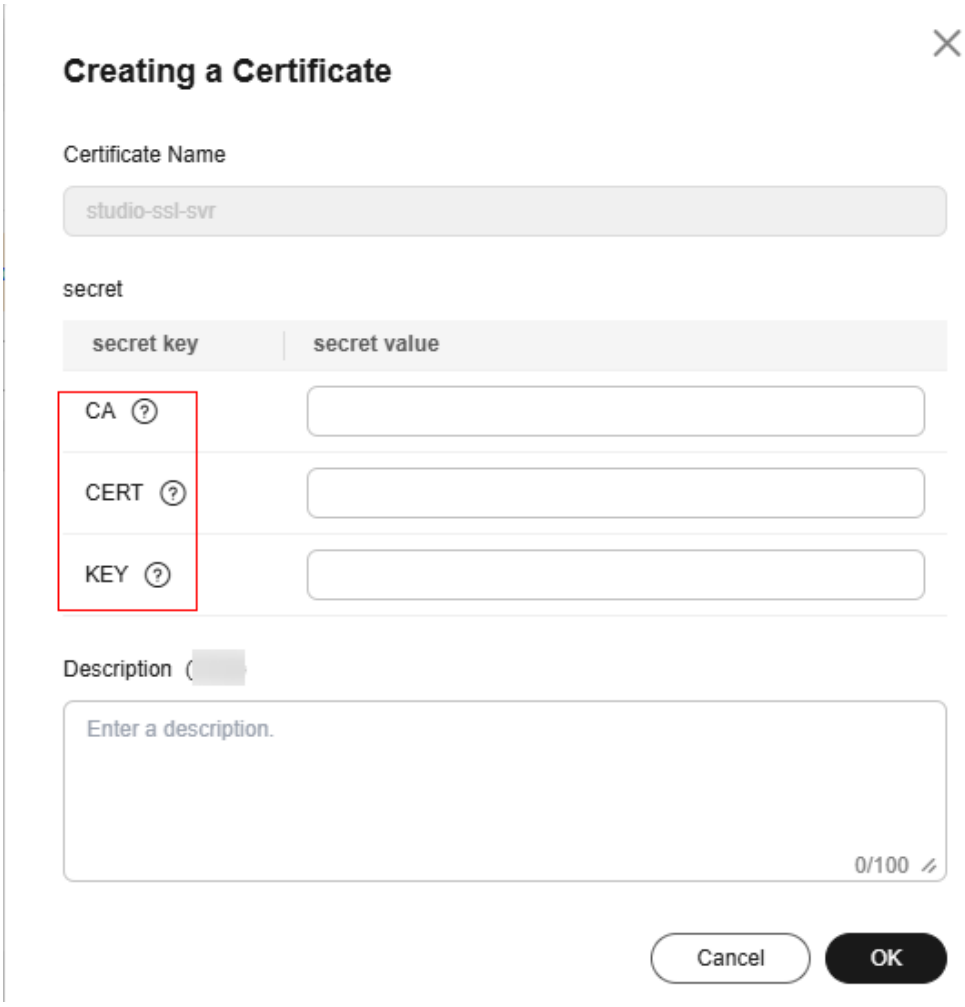
2. After the certificate is created, run the **ls** command to view the generated certificate file. The mapping between certificate files and certificate items on ModelArts Studio is as follows:

   ```
   ca.crt -- CA
   server.crt -- CERT
   server.key -- KEY
   ```

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3. On the ModelArts Studio homepage, click **Setting** in the upper right corner. On the **Resource Pool Management** > **Edge Resource Pools** tab page, click **Create Certificate** and set certificate parameters and description.

Figure 9-8 Certificate key



4. You can run the **view** command to view the certificate key value, for example, **view server.crt**. The key value entered on ModelArts Studio must be encrypted using Base64 (**echo -n** "Copied key value" **| base64**).

Step 5: Creating a Load Balancer

After an edge resource pool is created, go back to the homepage of ModelArts Studio and click **Setting** to create a load balancer and a monitoring plug-in for the edge resource pool.

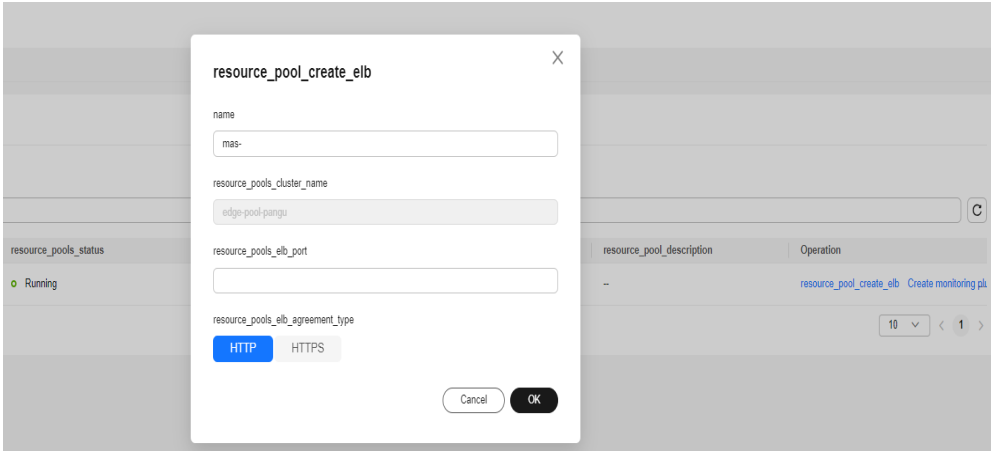
To create a load balancer, perform the following steps:

1. Log in to ModelArts Studio and click **Setting** in the upper right corner.
2. On the **Resource Pool Management** tab page, click the **Edge Resource Pool** tab. View the created edge resource pool, and click **resource_pool_create_elb** in the **Operation** column.

 Copyright © Huawei Cloud Computing Technologies Co., Ltd.

3. Enter the load balancer name (in the format of **mas-***xxx*), set the listening port (ranging from 30000 to 40000), and set the protocol type (using HTTP or HTTPS requests for calling the inference model). Click **OK**.

Figure 9-9 Creating a load balancer



 Copyright © Huawei Cloud Computing Technologies Co., Ltd.