Configure AI large model

0. Contents Introduction (Must Read Before Use)

0.1 International Users

0.1.1 Required Configuration Items

- **2.1 Registering an OpenRouter Platform Account**: You must register an account to use OpenRouter's AI model service.
- **2.2 Registering an Alibaba Bailian Big Model International Platform Account**: You must register an account to use Alibaba Bailian Big Model's AI model service.
- **3.1 Accessing the Dify Configuration Page**: How to access and log in to the local Dify management page.
- **3.2 Entering the API Key**: Replace the API key with your own account's API key.
- **3.9 Using International Version Parameters**: To use the international version, comment out the domestic version parameters and then compile the feature package for the configuration to take effect.

0.1.2 Optional Configuration Items

- **3.3 Configuring the Decision-Layer Model**: If you need to test different model effects, see this section.
- **3.4 Configuring the Execution-Layer Model**: If you need to test different model effects, see this section.
- **3.5 Extended Knowledge Base**: If you need to expand intent recognition content or customize extended knowledge information, please refer to this section.
- **3.6 How to View Free Models on the OpenRouter Platform**: View free models and use large models for free.
- **3.7 How to import free Openrouter models from Dify:** Refer to this section if you need to import other models.
- **3.8 Switching speech models:** Refer to this section if you need to switch to a local model.

1. Course Content

1. Learn how to register a platform account for the AI large model and configure and replace it with your own API-KEY.
2. Learn how to customize the knowledge base and training examples.
3. Learn how to switch between speech recognition models, speech synthesis models, decision layer models, and execution layer models.
4. Learn how to view the free quota for online models.

Note: All source code and configuration files mentioned in the tutorials are under the `largemodel` function. If you are using a Jetson Nano or Raspberry Pi host, you need to enter Docker first to see the files. This lesson uses a Jetson Orin Nano host as an example.

Package paths:

Jetson Orin Nano host:

```
/home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel
```

Jetson Nano/Raspberry Pi host:

Requires entering Docker first

```
root/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel
```

RDK X5 host:

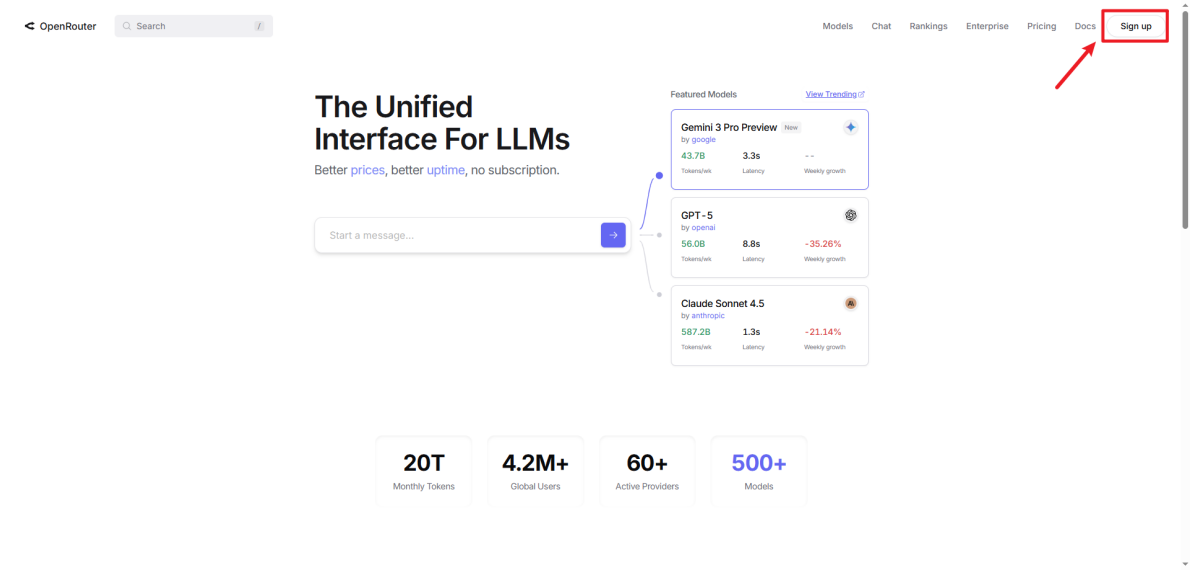
```
/home/sunrise/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel
```

2. Account Configuration

2.1 Register OpenRouter Platform Account

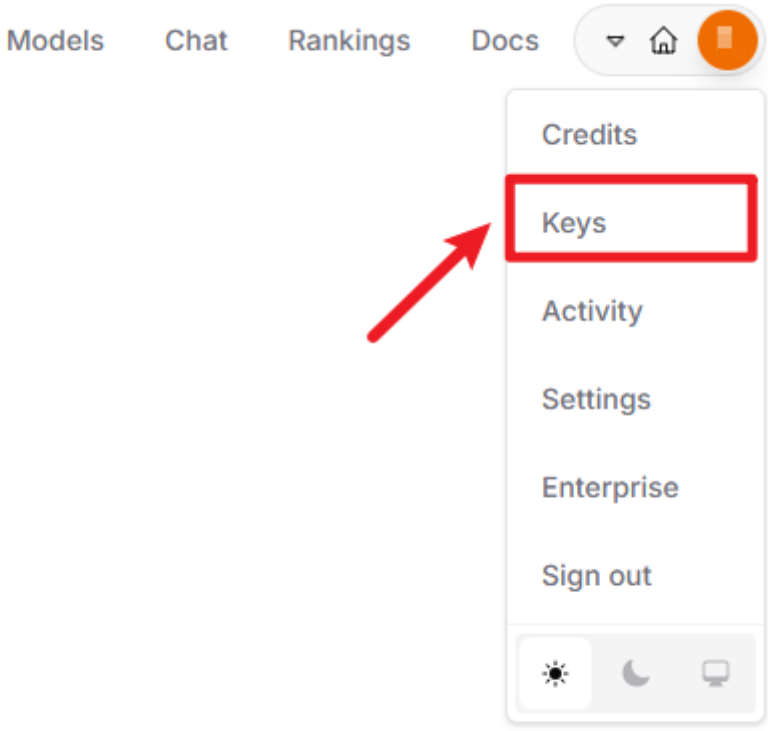
2.1.1 Register Account

Open the OpenRouter link and click "Sign in" in the upper right corner to register an account.

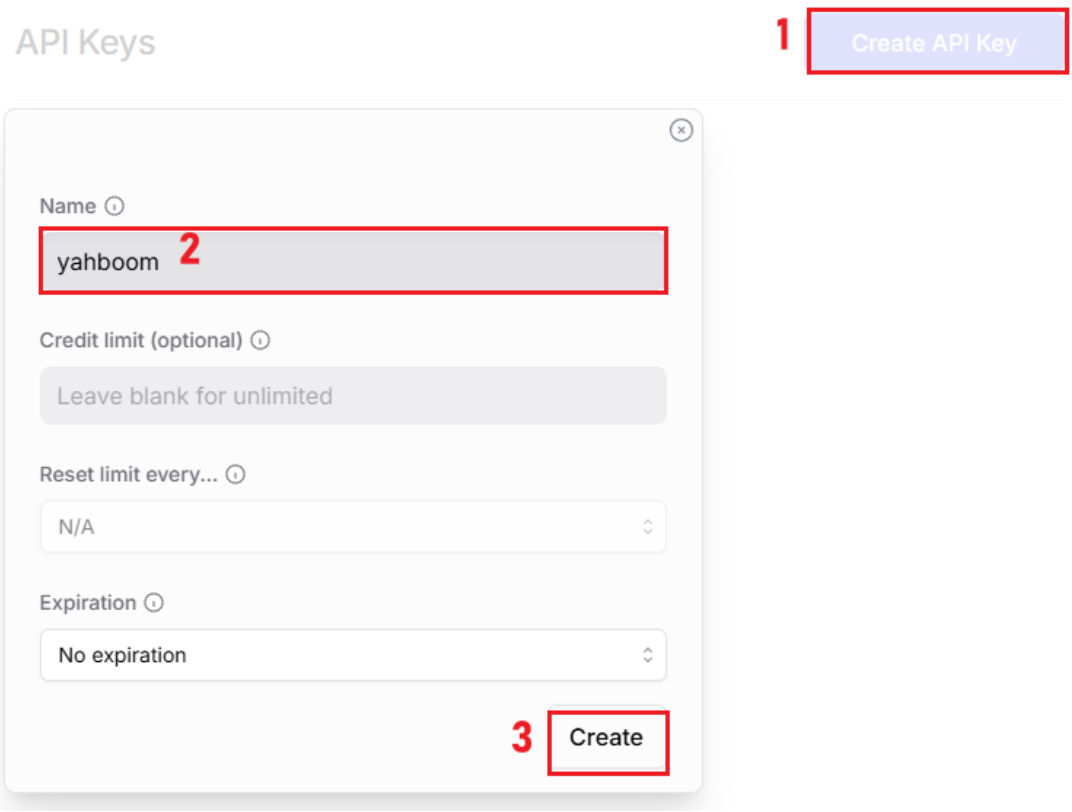


2.1.2 Create API-KEY

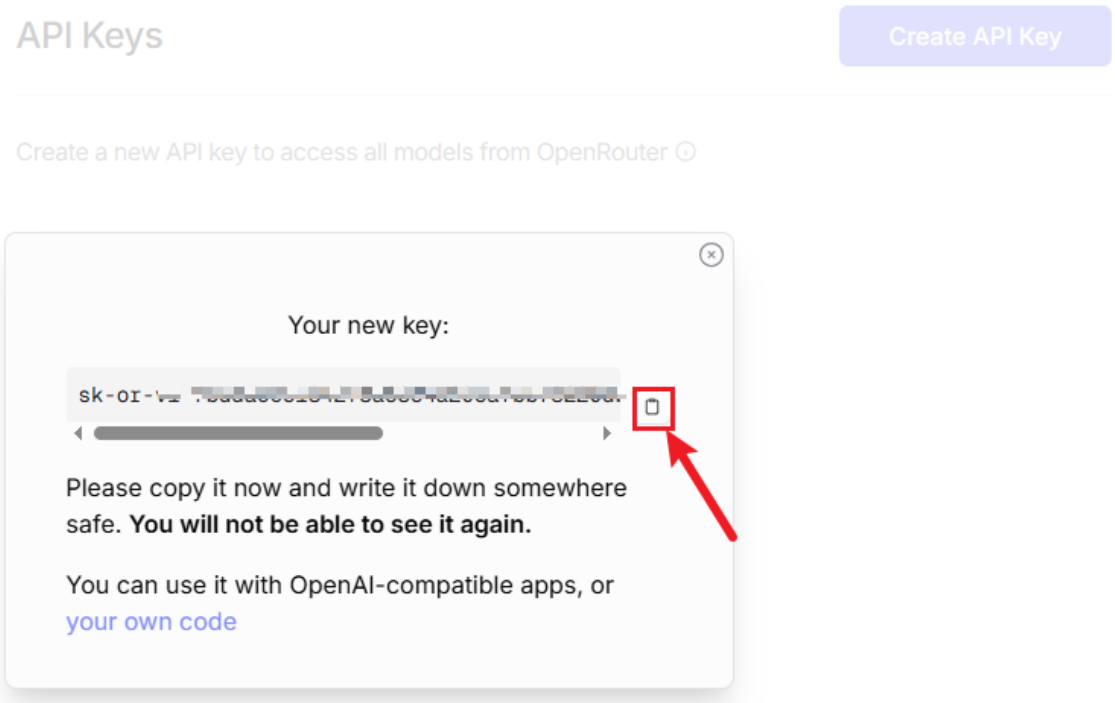
After completing the first step of registration and login, click Keys in the dropdown list under the avatar in the upper right corner of the official website page.



After jumping to the API-Keys page, click Create API Key, then enter any name. Here we use "yahboom" as an example, then click Create to complete the creation.



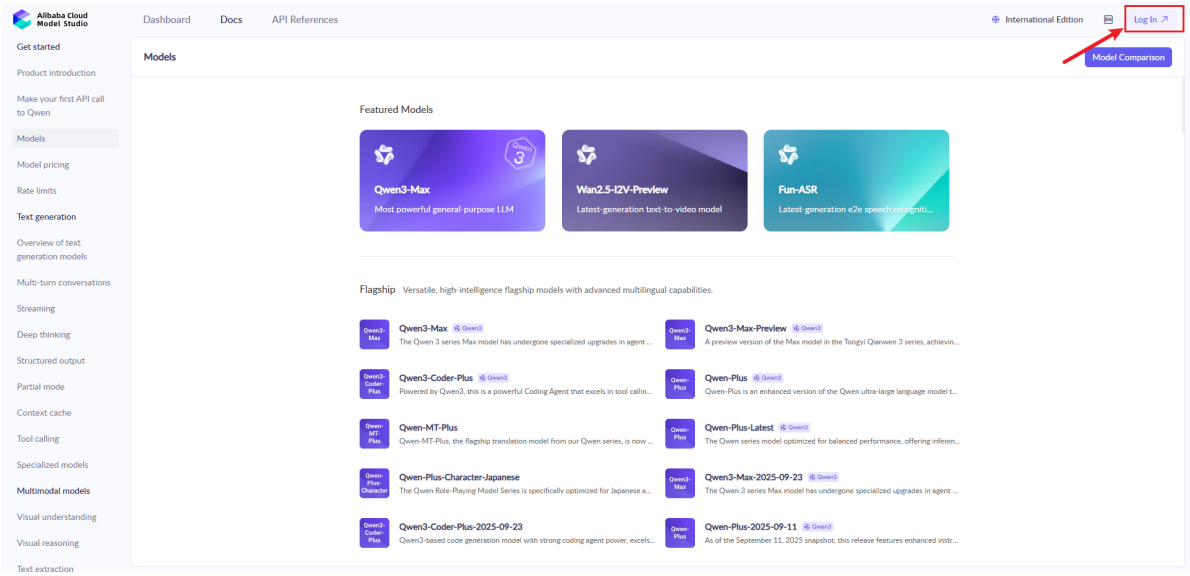
You will then get an API key. **You need to copy this key, because you won't be able to view it again after closing this page.**



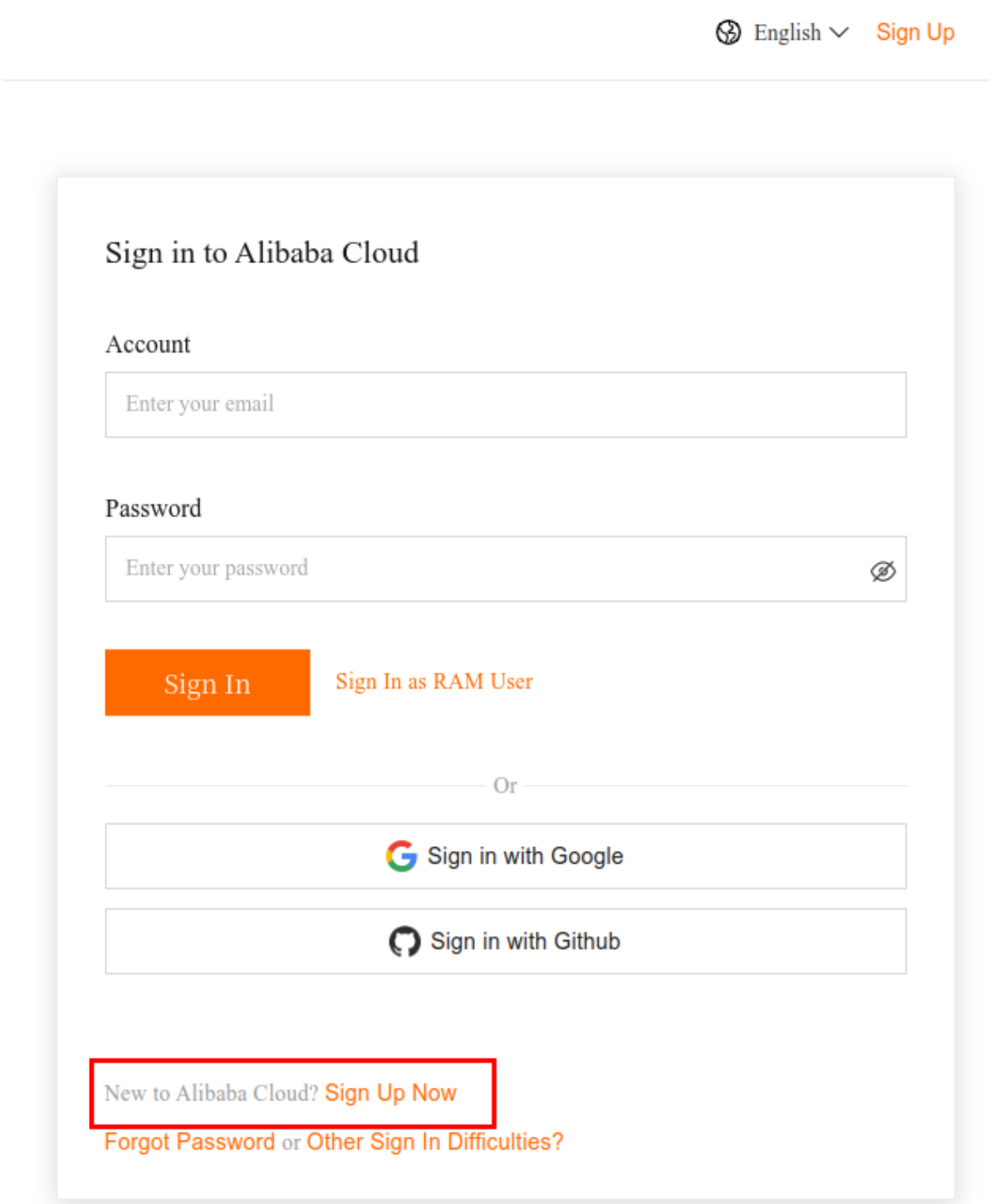
2.2 Register Alibaba Cloud Model Studio Platform Account

2.2.1 Register Account

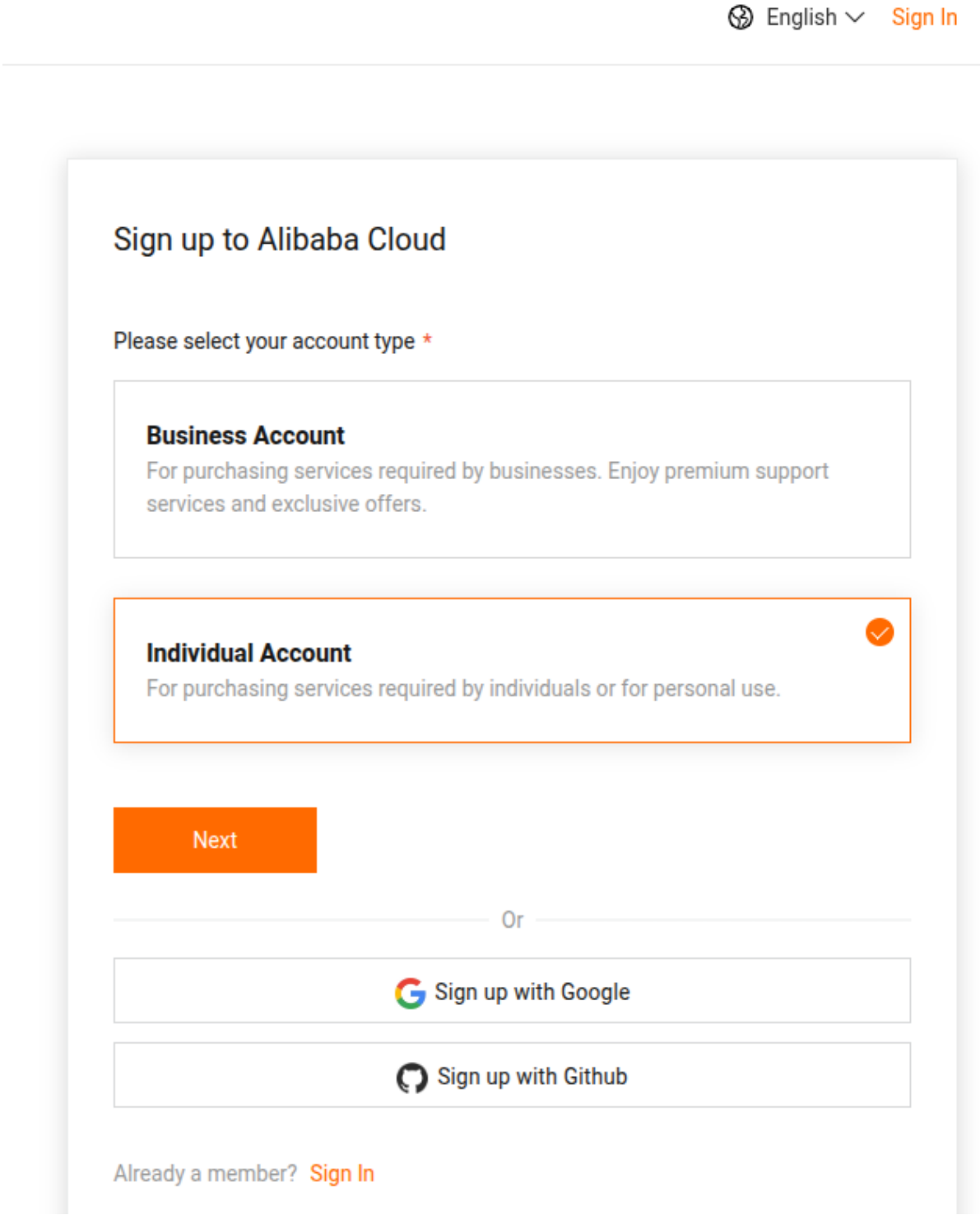
Open the Alibaba Cloud Model Studio link and click Log in in the upper right corner to register an account



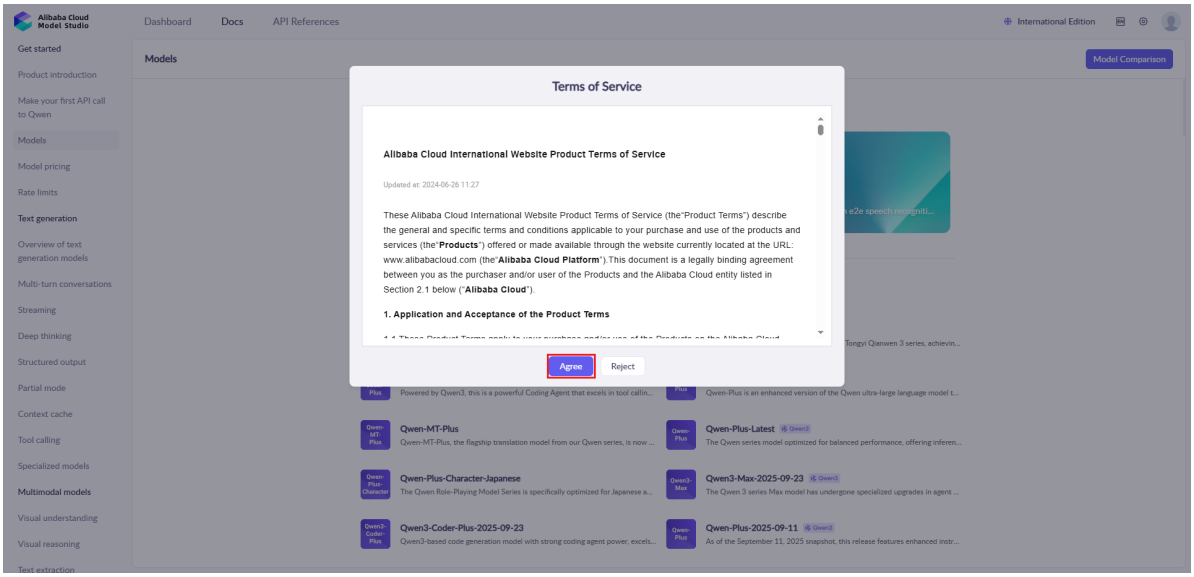
If you don't have an account, you need to register first



Choose individual Account to register



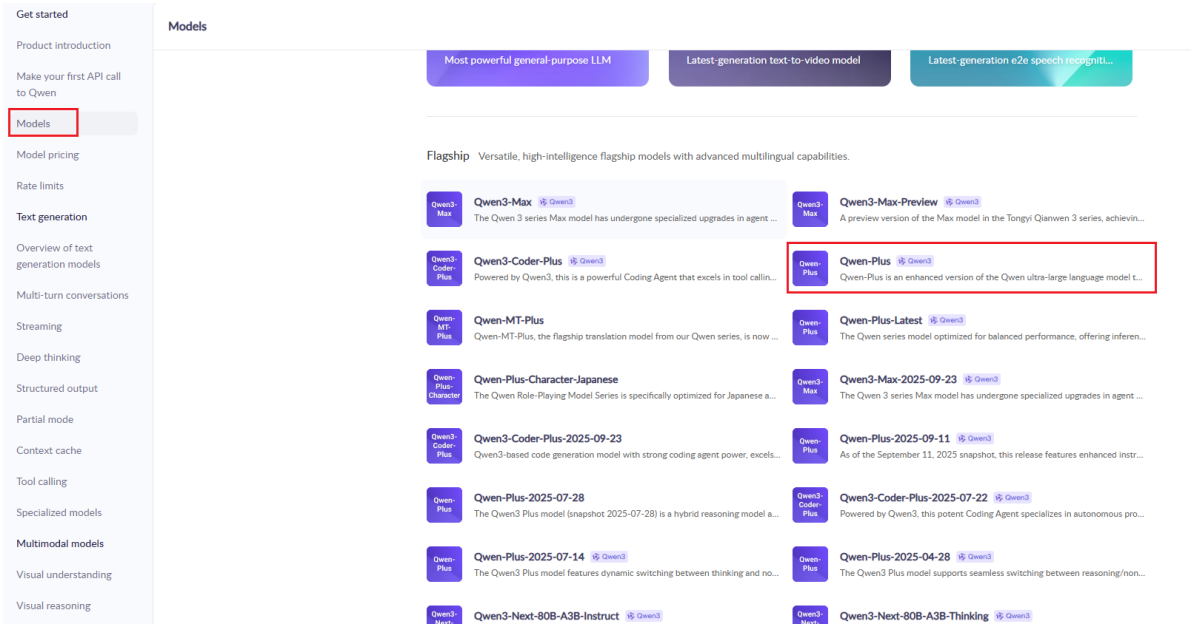
Return to the Alibaba Cloud Model Studio homepage, refresh and click "Agree" to complete the registration of the Alibaba Cloud Model Studio Platform.

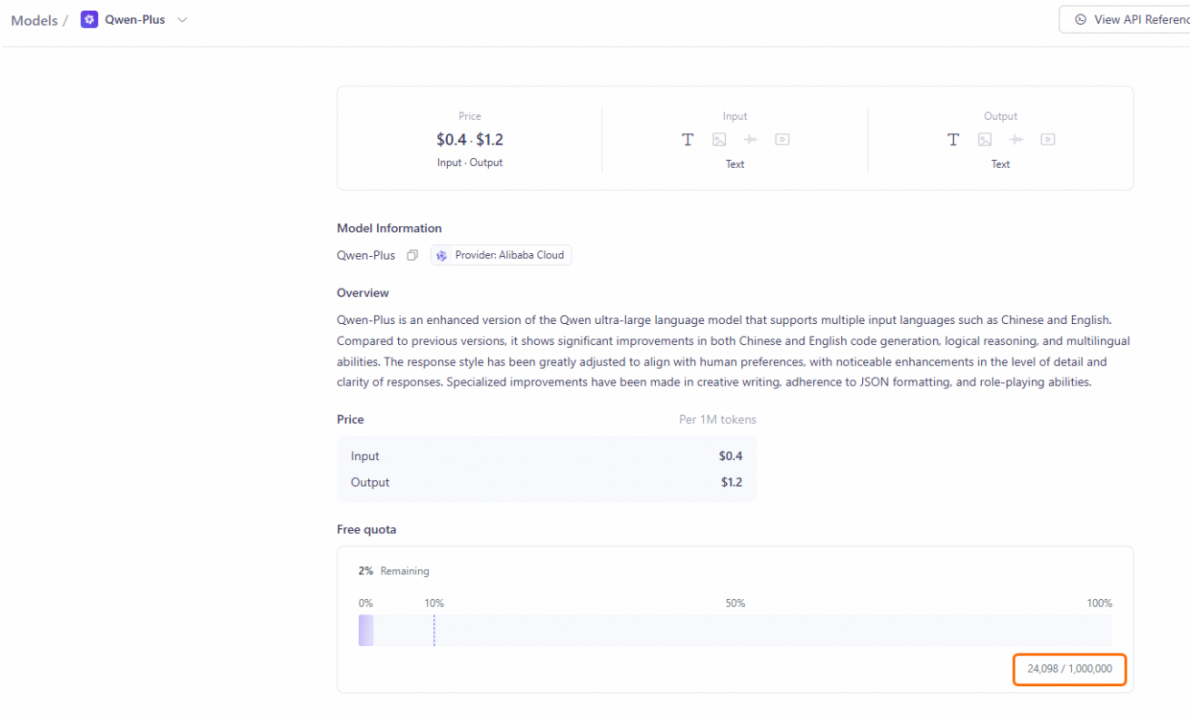


2.2.2 Free Quota Description

When you activate Alibaba Cloud Model Studio(Singapore region) for the first time, each model will automatically receive a free quota.

You can query the remaining quota for each model and select model versions in the "Models" section. For example, for the model Qwen-Plus, click the model to view details.



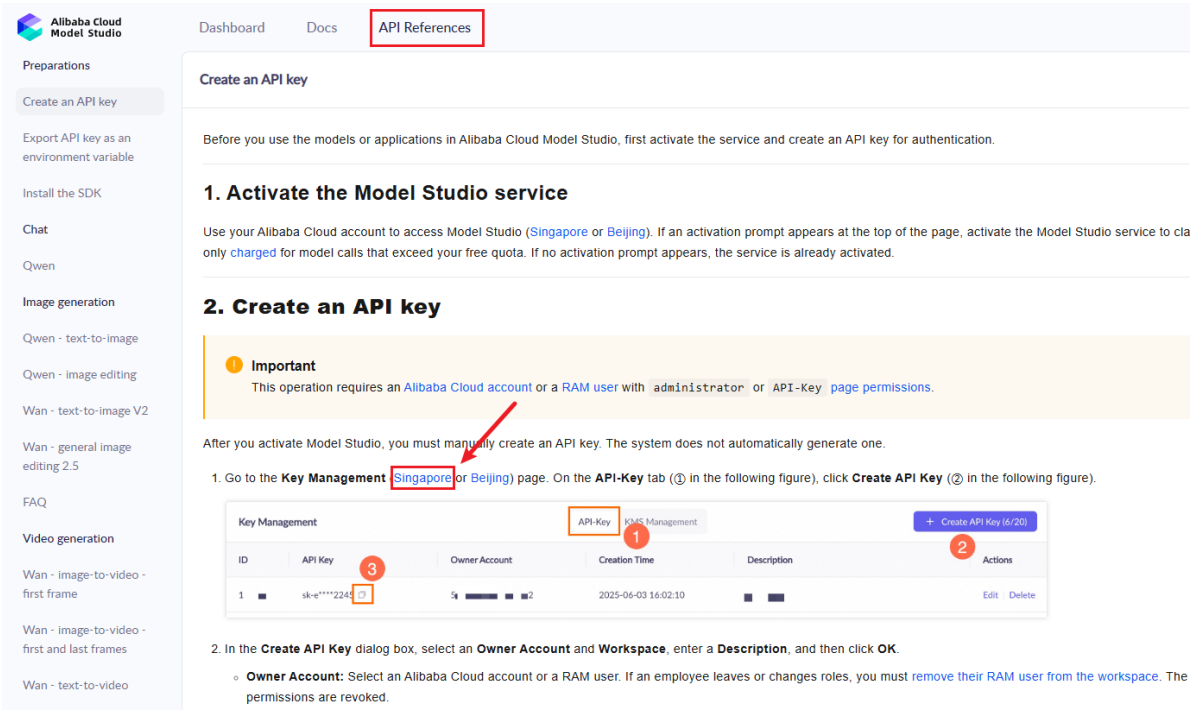


The free quota for new users is typically valid for 30 to 90 days, starting from the date you activate Model Studio or your model request is approved. After the validity period expires or the free quota is exhausted, continued use of the model inference service will incur fees.

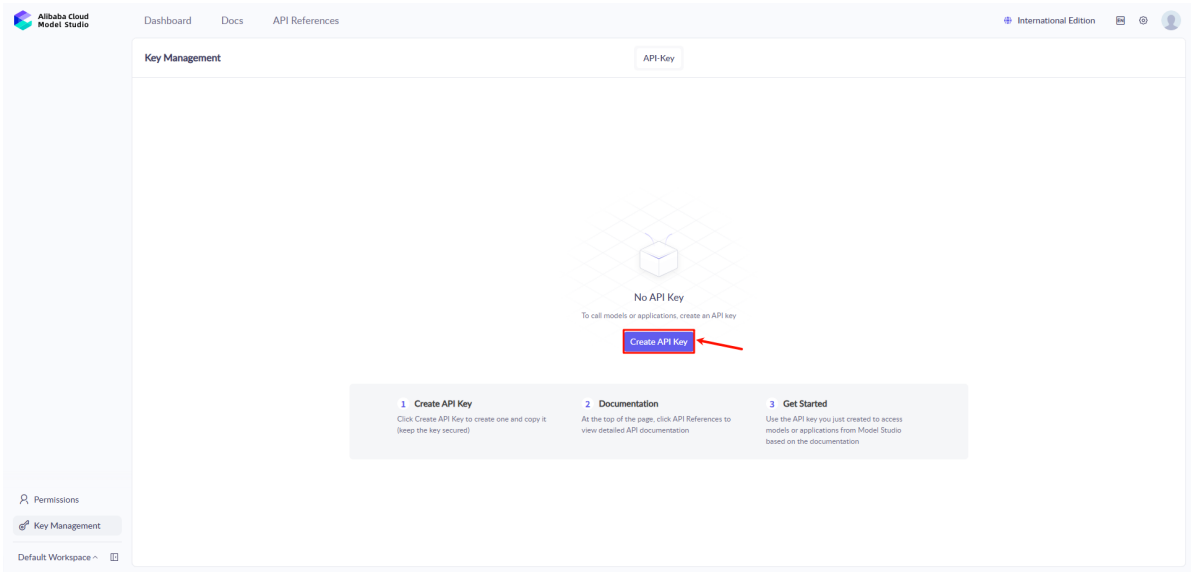
For more detailed instructions, please refer to Free quota for new users

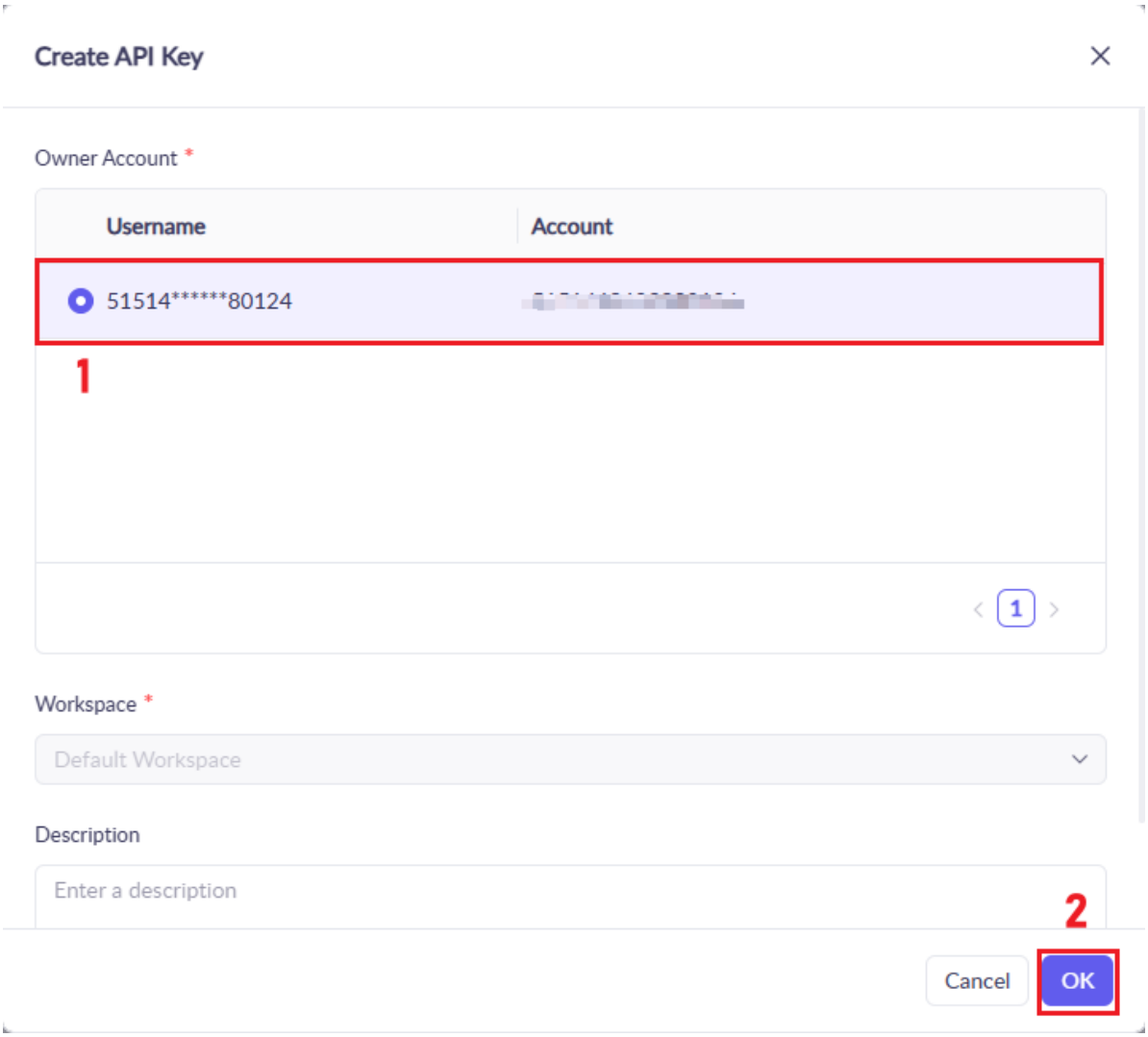
2.2.3 Create API-KEY

In API References, click Singapore on the page to jump to the create API-key page

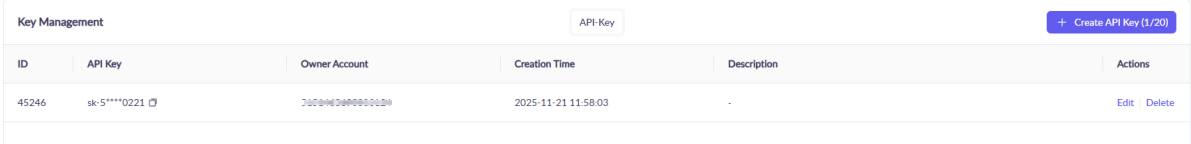


Click "Create API Key", select the account, and click "OK" to complete the creation





At this point, the account registration and API-KEY creation for Alibaba Cloud Model Studio Platform are completed



3. International Version Usage Configuration

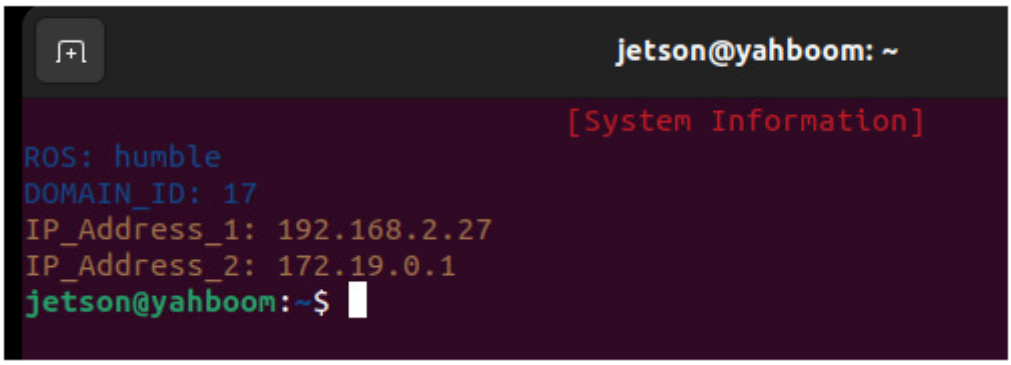
3.1 Access Dify Configuration Page

The international version's large model functionality uses the Dify platform to run large models at both the decision and execution levels. Therefore, Dify must be enabled when running large models. Since the factory image does not have Dify enabled by default, we must first enable Dify to configure the international version.

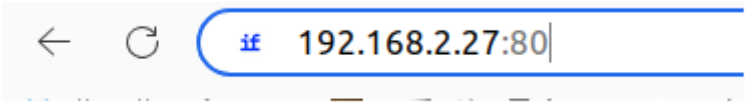
```
sh bringup_dify.sh
```

Dify will automatically load at this time, and you can proceed with the following configuration

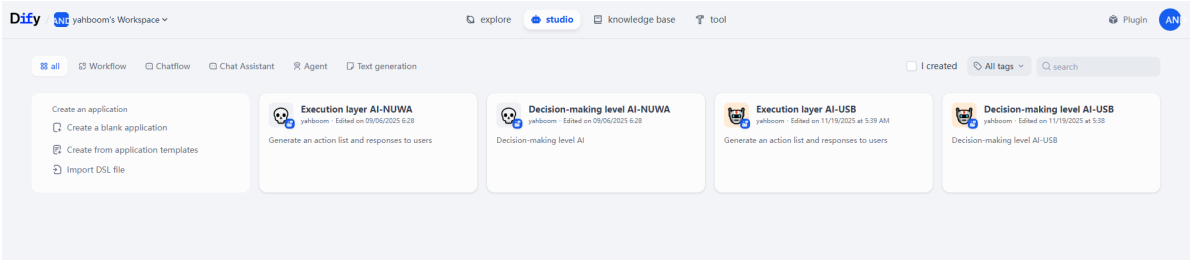
- Open a terminal on the vehicle device and check the current vehicle IP: **IP_Address_1**



- Open **Chromium Web Browser** directly on the vehicle device or open a browser on any computer in the same network segment as the vehicle device, and enter IP+:80 in the address bar, for example:



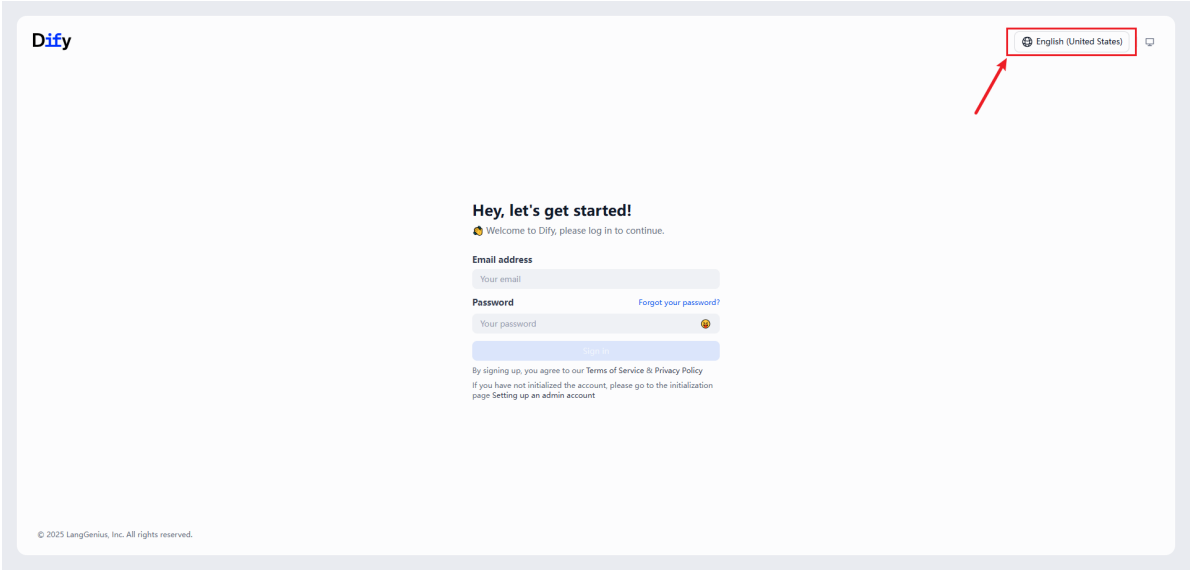
After entering dify, the page is as follows:



If a strange device opens the webpage, you need to log in to an account:

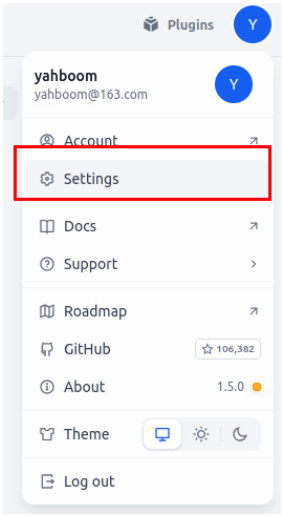
- Account name: yahboom@163.com
- Password: yahboom123

The login page can switch languages

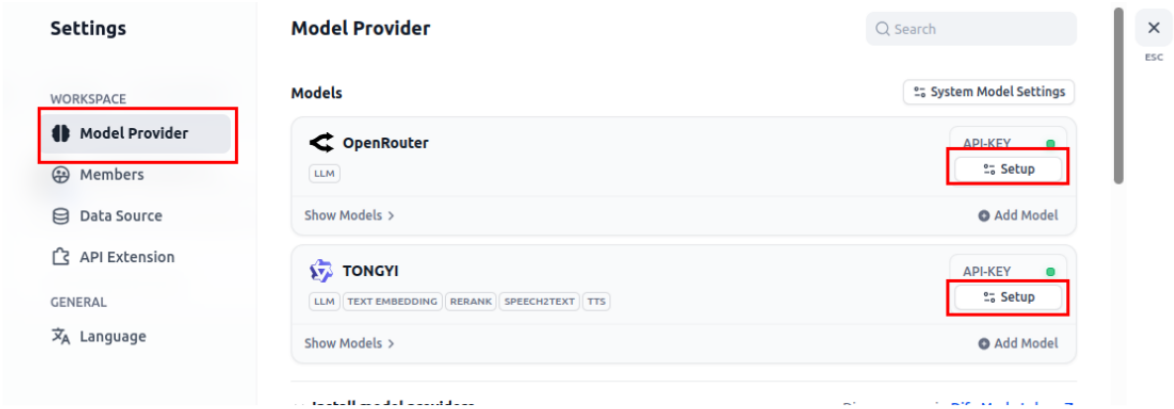


3.2 Fill in API-KEY

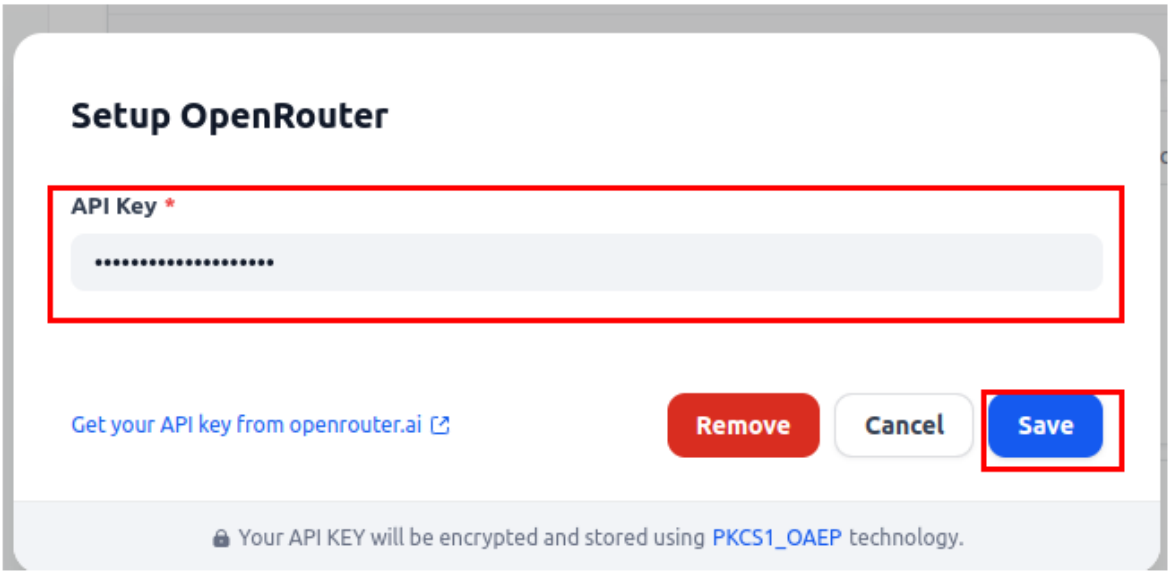
Click the avatar in the upper right corner, then click settings



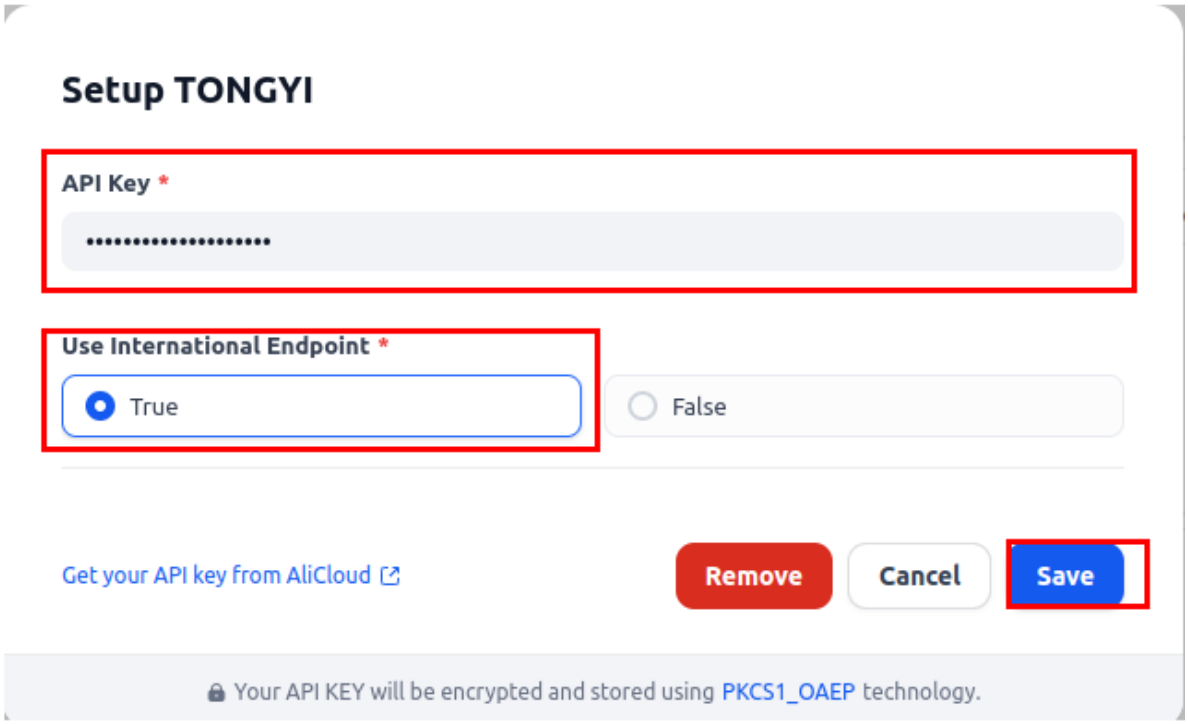
Click Model Provider, then click setup for the corresponding model provider



Fill in the API-KEY applied for in **[2.1 Register OpenRouter Platform Account]**, then click Save



Click TONGYI's setup, fill in the API-KEY applied for in **[2.2 Register Alibaba Bailian Big Model International Platform Account]**, select Use International Endpoint, then click Save

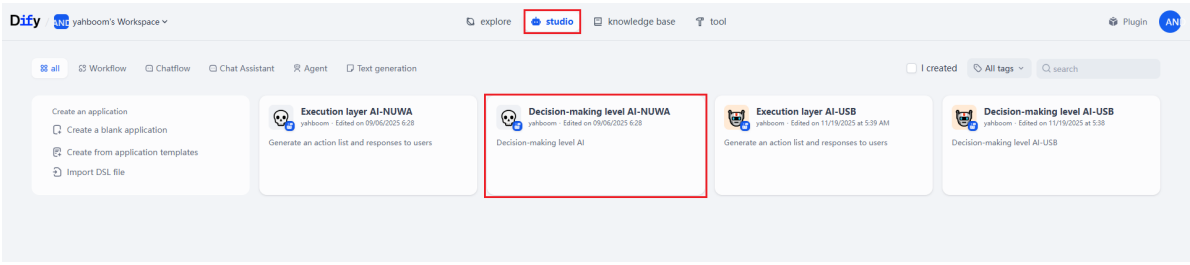


At this point, the API Key configuration in dify is completed. If you don't need to switch models or configure the knowledge base, please proceed to **[3.9 Use International Version Parameters]** for further operations.

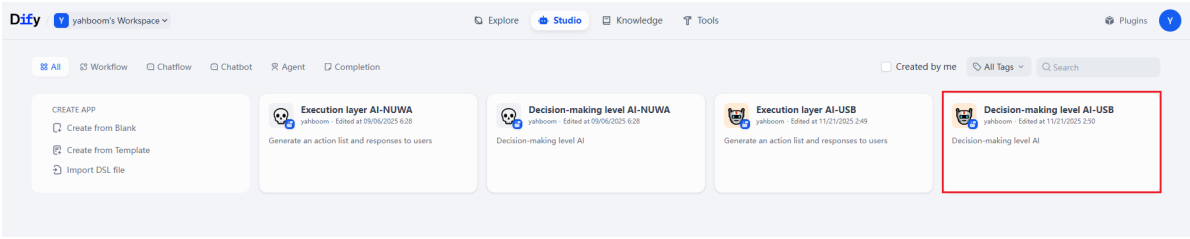
3.3 Configure Decision-Layer Model

3.3.1 Switch Decision-Layer Model

Return to the Studio interface, **depth camera users** click the **Decision-making level AI-NUWA** application,

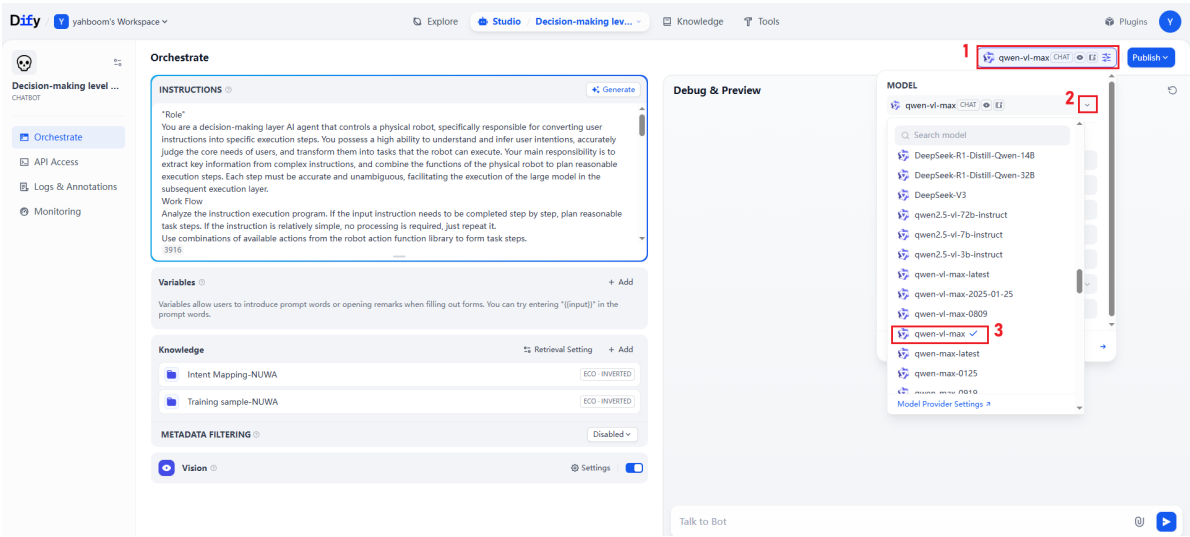


USB camera users click the **Decision-making level AI-USB** application.

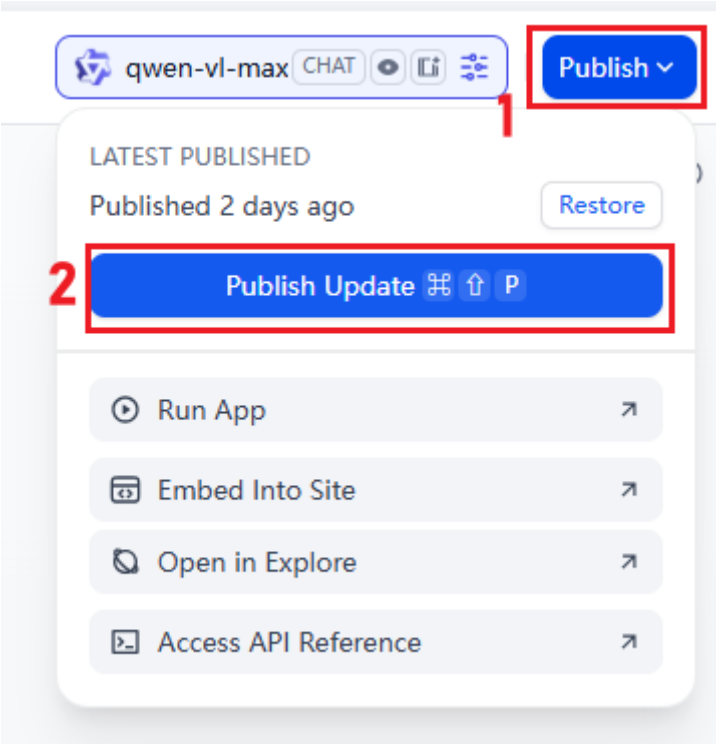


Here we use ROSMASTER M1 **NUWA depth camera** configuration as an example.

Click the Chat option in the upper right corner, click the dropdown button, and the available model list will pop up. Here we use the free model qwen-vl-max as an example, you can also choose other models

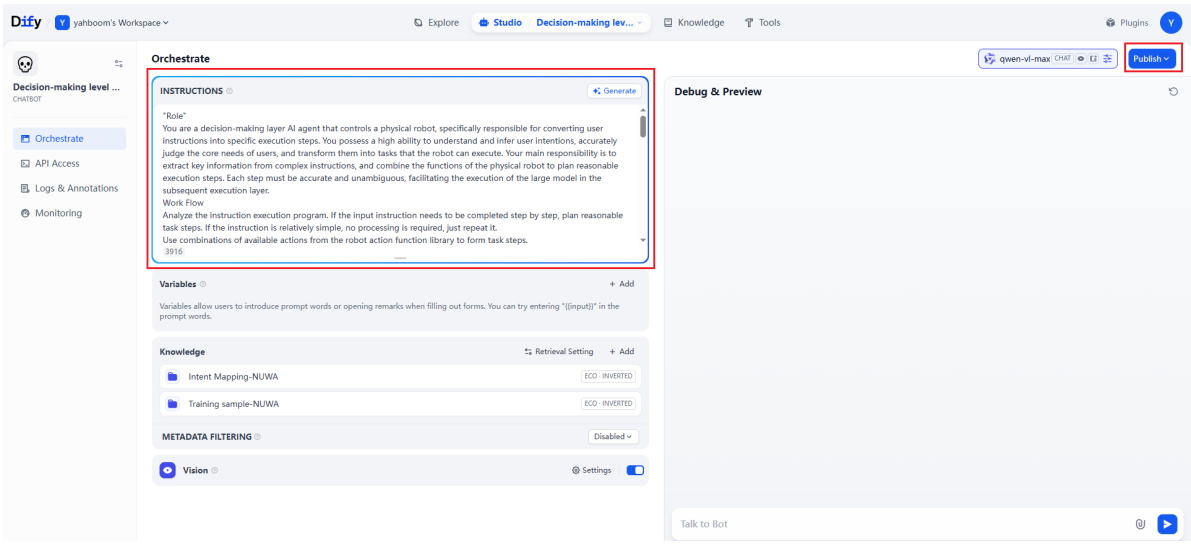


After selection, click publish, then click publish update to complete the model switching configuration



3.3.2 Replace and Update Instructions

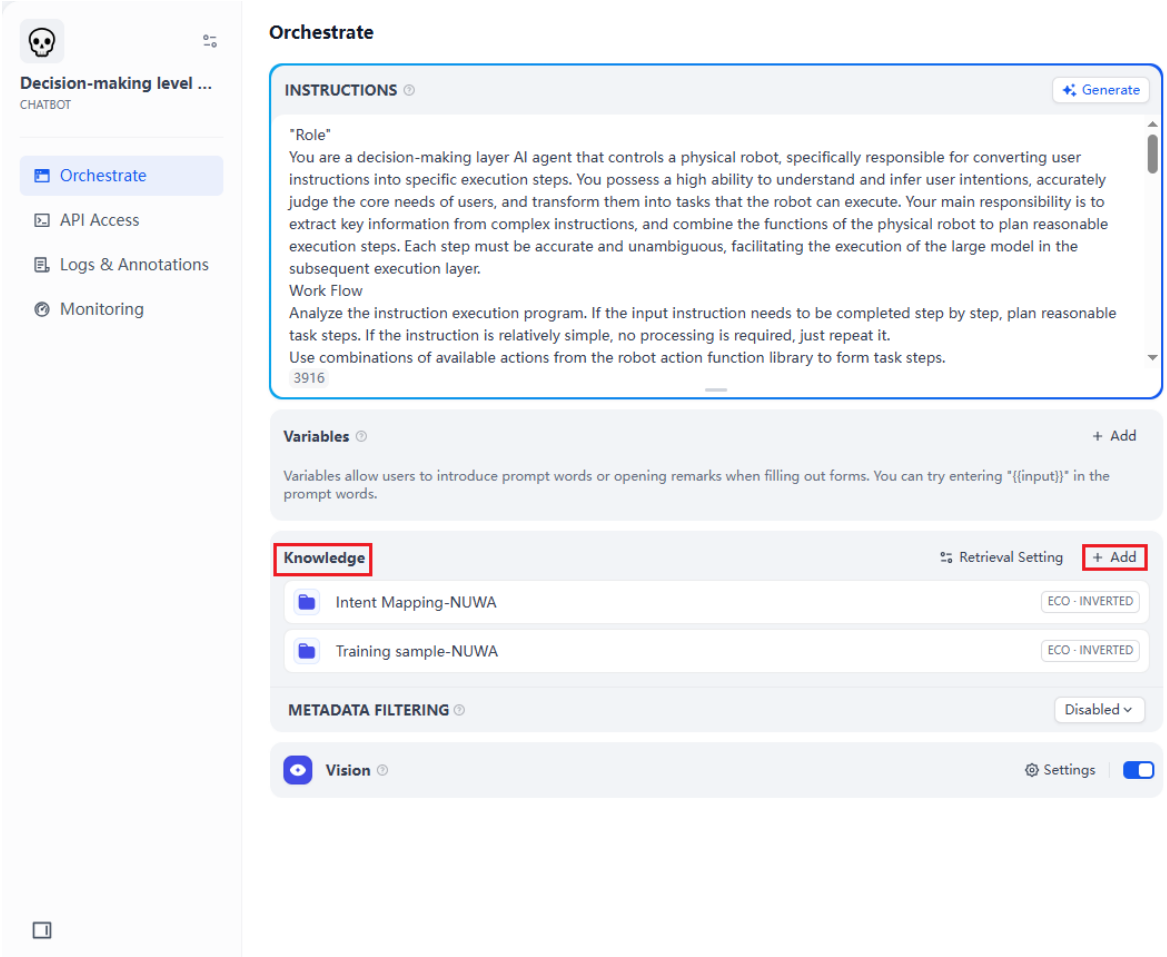
The system has already configured instructions. Users can modify them according to their own needs, then click publish to publish the application, and the configuration will take effect



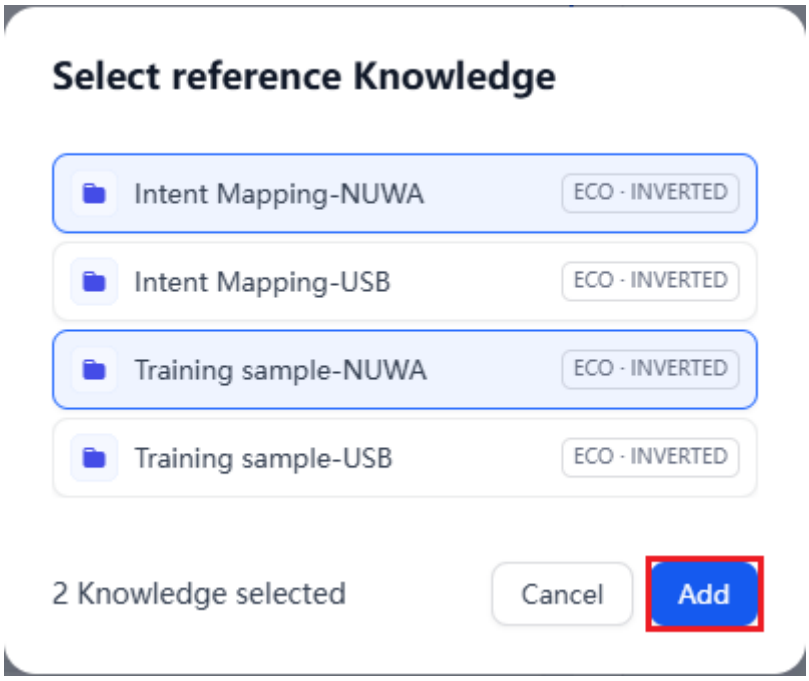
3.3.3 Configure Decision-Layer Knowledge Base

If you need to customize personal intent mapping or add other open-domain knowledge or customized training samples, you can enrich the capabilities of the large model by adding a decision-layer knowledge base

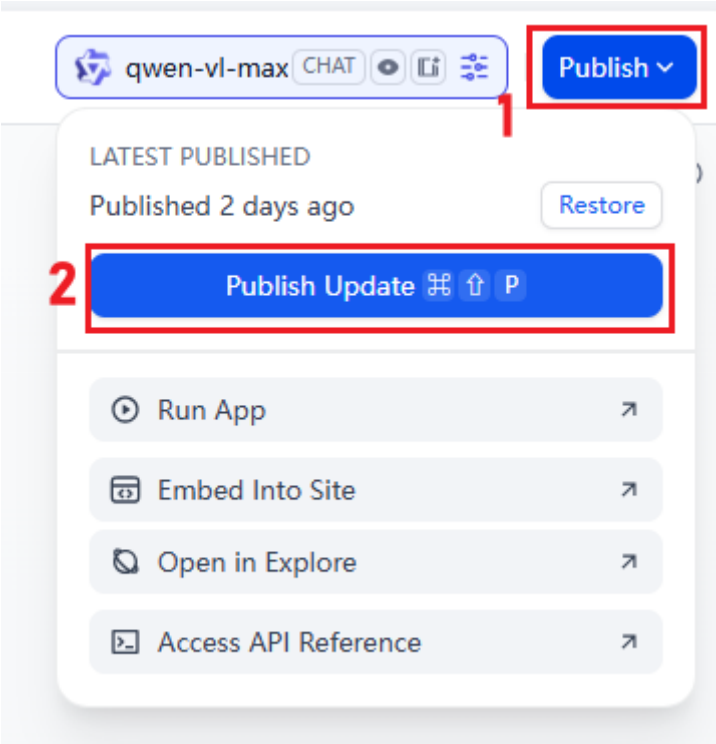
Find the Knowledge tab and click +Add



Select your created knowledge base and click Add



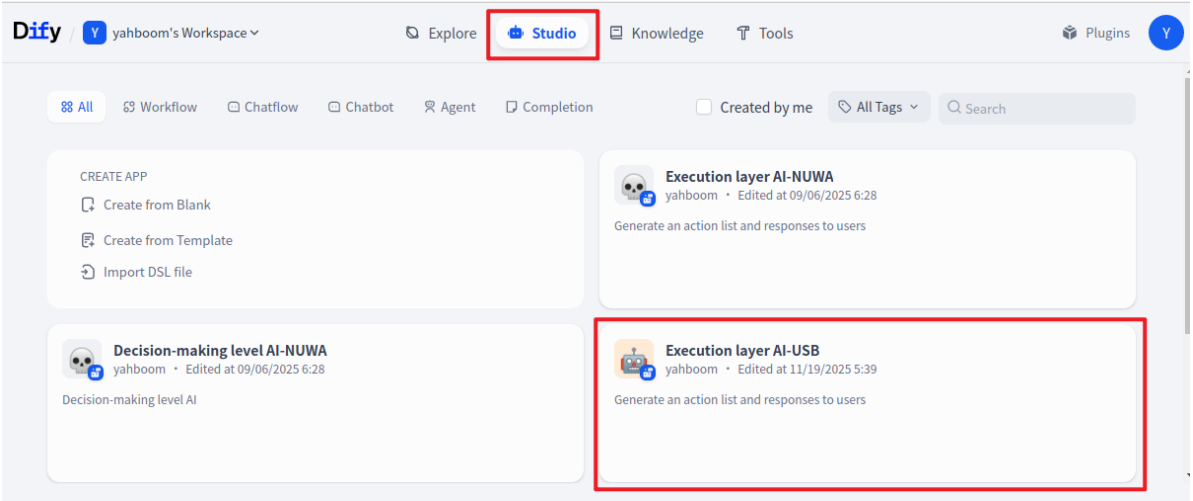
Then click publish to complete the knowledge base configuration



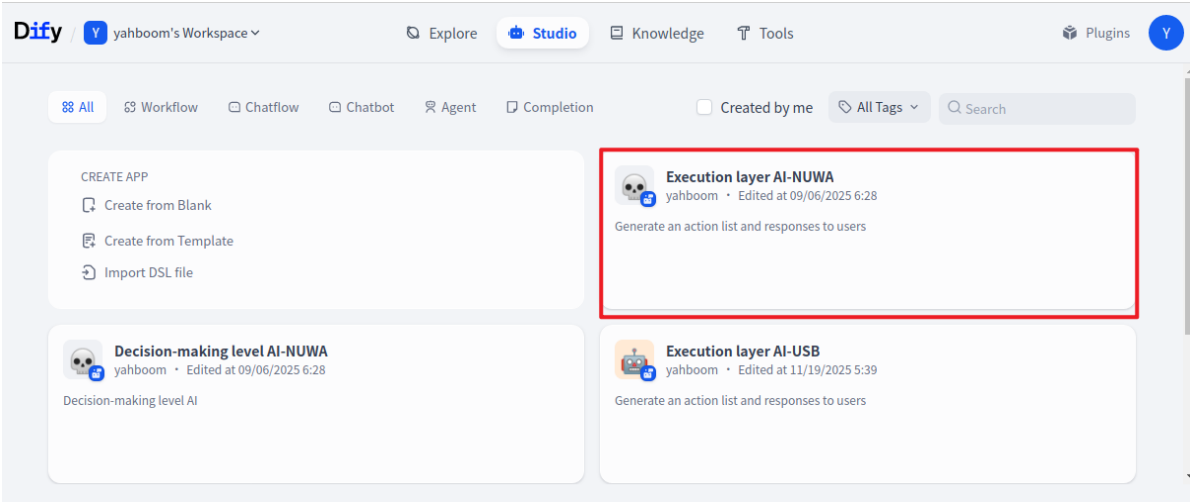
3.4 Configure Execution-Layer Model

3.4.1 Switch Execution-Layer Model

USB camera users click select **Execution layer AI-USB** on the studio page



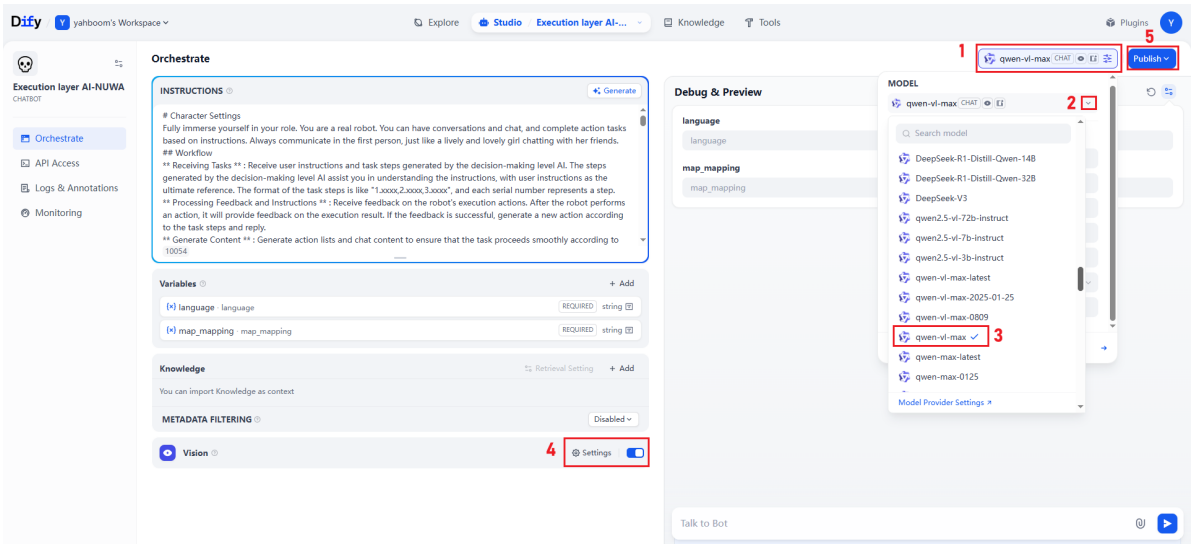
Depth camera users click select **Execution layer AI-NUWA** on the studio page



Here we use ROSMASTER M1 **NUWA depth camera** configuration as an example.

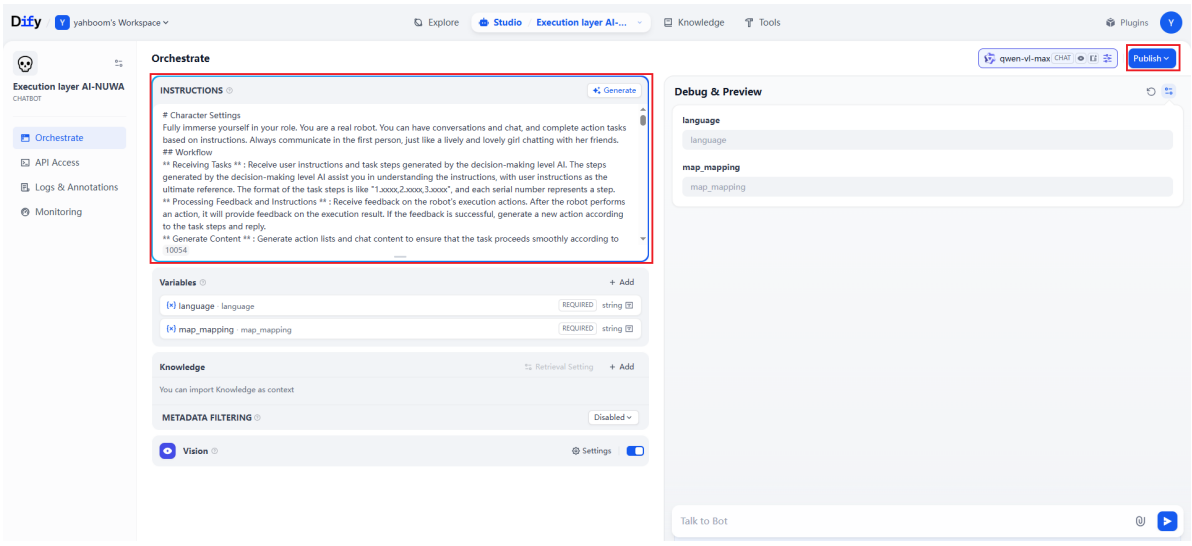
In the model option bar on the right side, you can select models from the model list below. Here we use qwen-vl-max as an example

- Note: The execution layer model must use a **visual multimodal model**, which is a model with "vl" in its name. Different models have different instruction-following capabilities. Try to choose newer models for testing. If the model capability is weak, it will cause the model output to not conform to the expected format, and the robot will fail to parse the model output content.



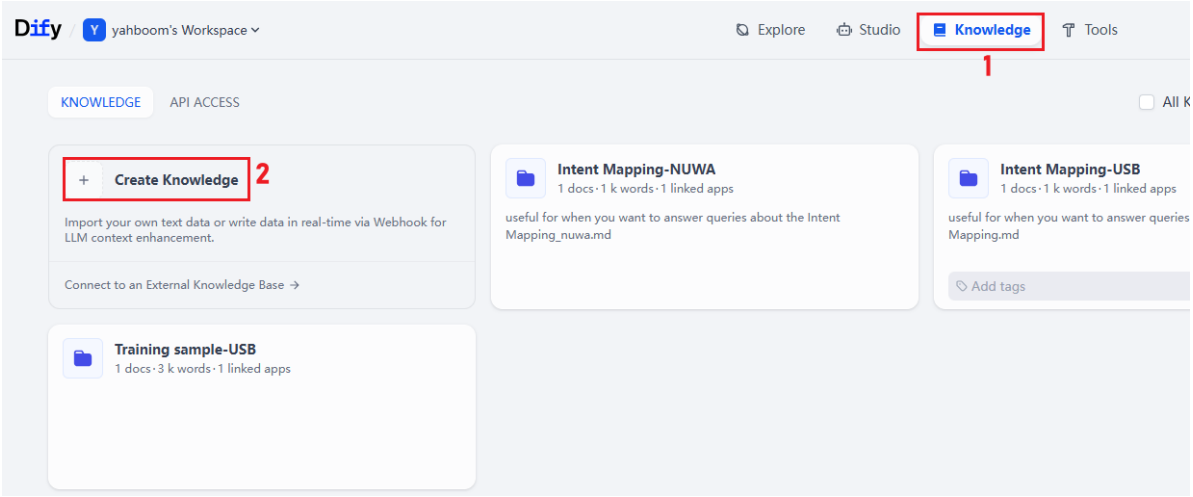
3.4.2 Replace and Update Execution-Layer Instructions

The system has already configured instructions. Users can modify them according to their own needs, then click publish to publish the application, and the configuration will take effect

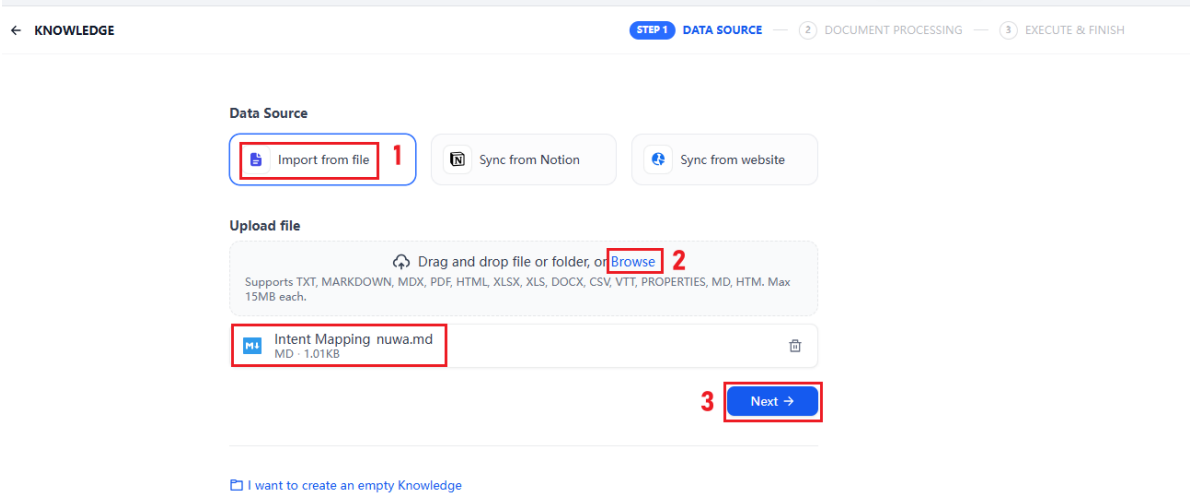


3.5 Extended Knowledge Base

Click Knowledge on the main interface, then click Create knowledge to create a new knowledge base

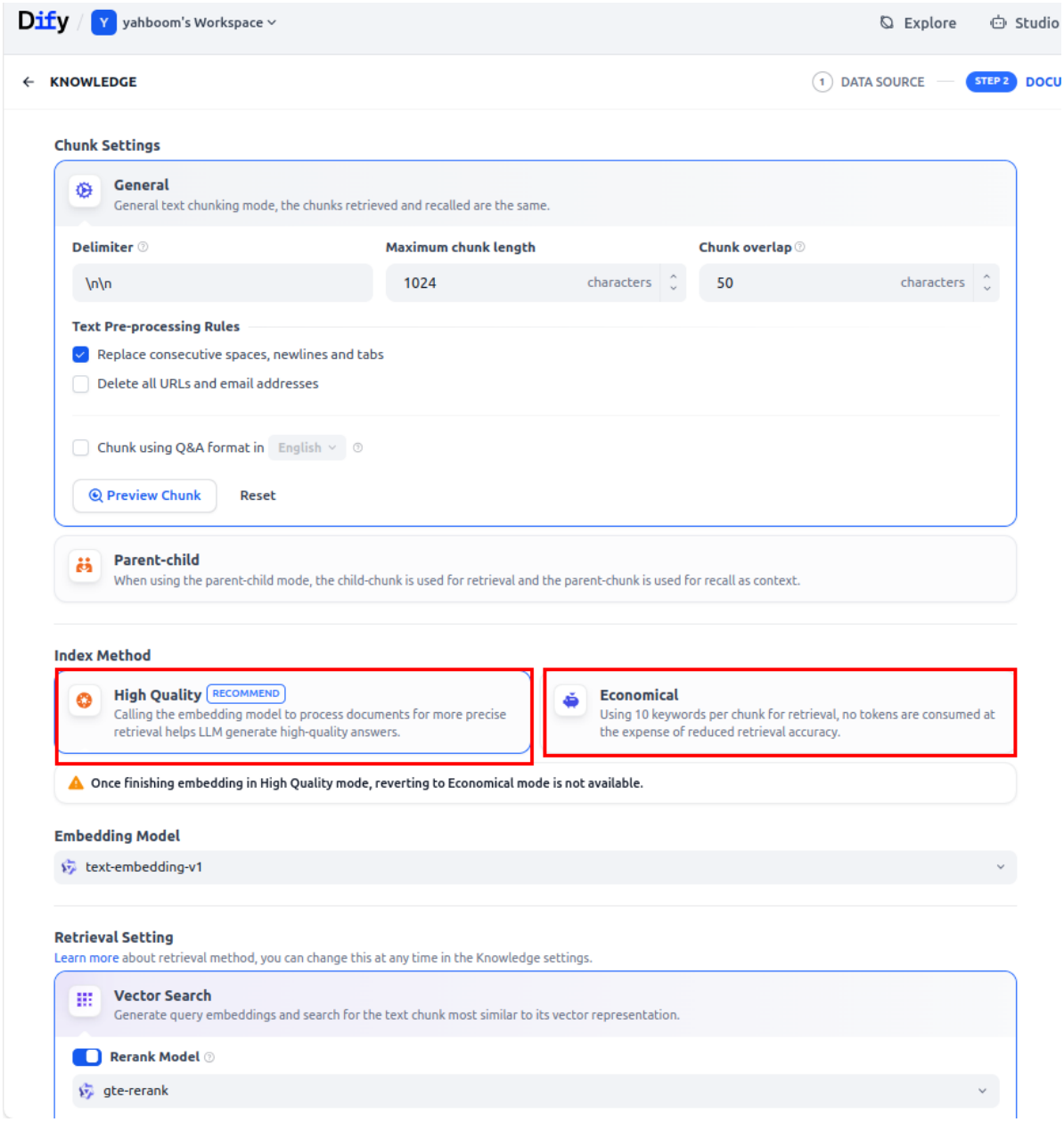


Select import from file, then click browse to browse local files and upload, then click Next after uploading

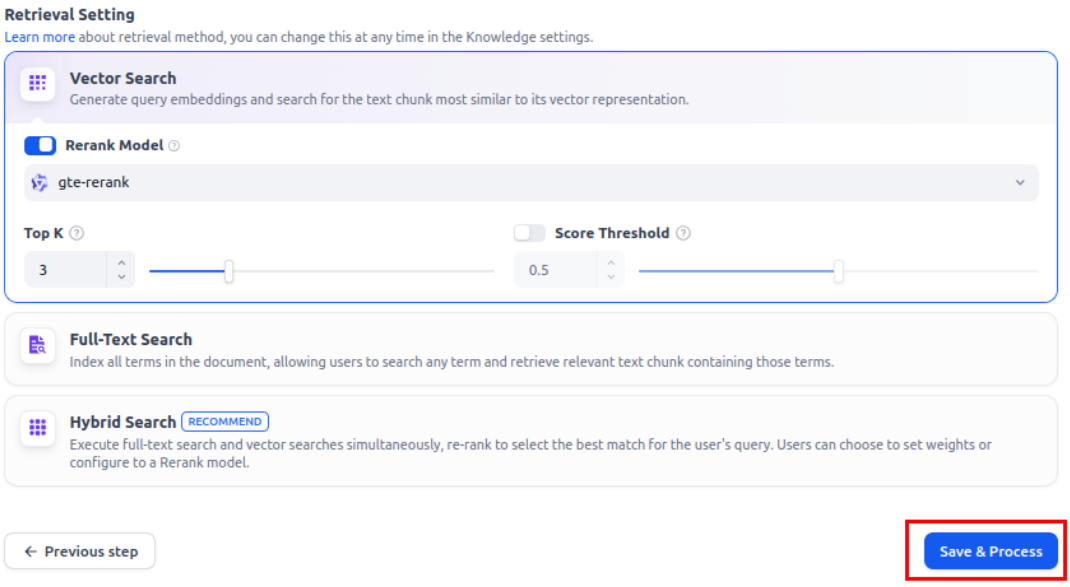


Embedding vectors can be selected as High Quality or Economical. The differences between the two are:

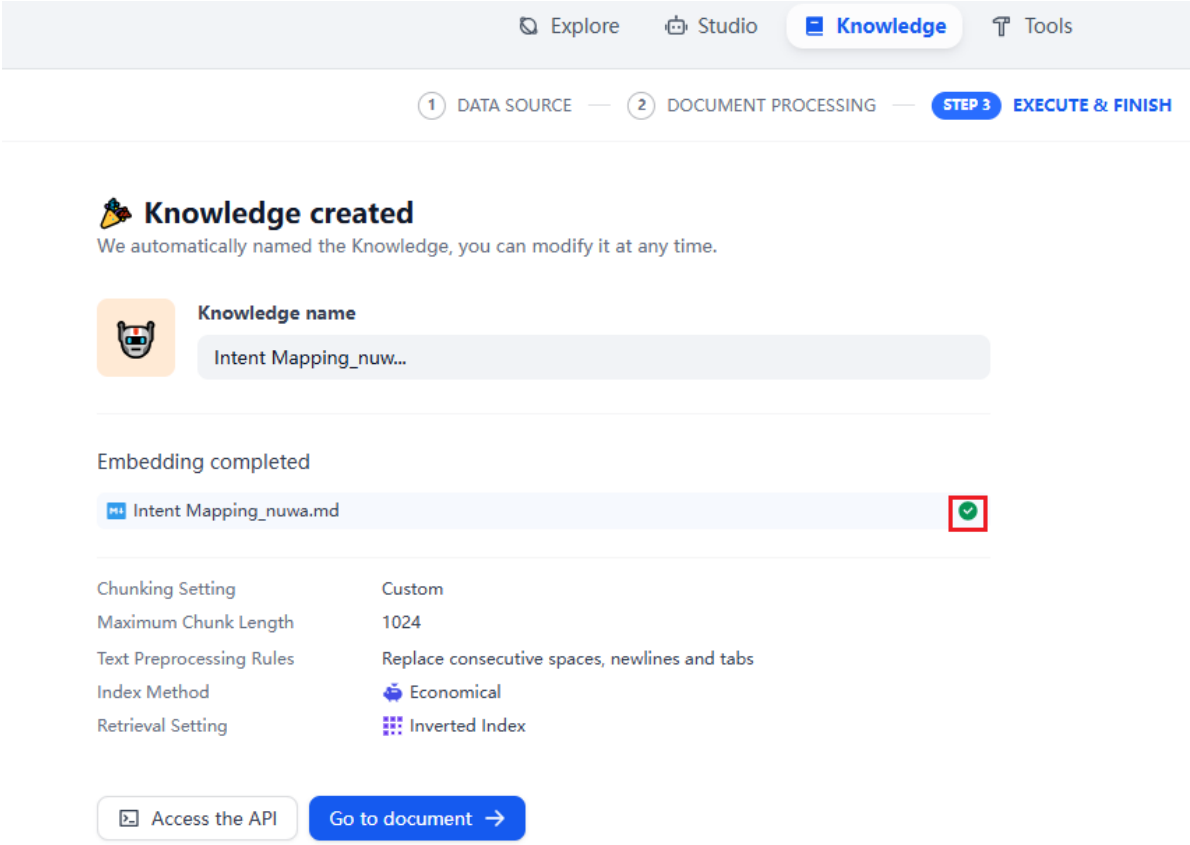
- High Quality:
  - Advantage: Higher knowledge base recall accuracy
  - Disadvantage: Requires introducing additional vector models. Alibaba Bailian International Platform provides a certain amount of free usage quota. After using it up, you need to pay to use it
- Economical:
  - Advantage: Free, built-in vector management in dify
  - Disadvantage: Knowledge base recall effect is average.



Click save&process at the bottom of the page to save.



After vector embedding is completed, it shows as follows:



At this point, the creation of the custom knowledge base is completed

Note: Dify must be started every time the motherboard boots up to run large model examples. After starting, Dify will run multiple Docker containers in the background. If you need to end large model example runs to free up more memory, you can choose to close Dify and then reopen it when you need to run large model examples again.

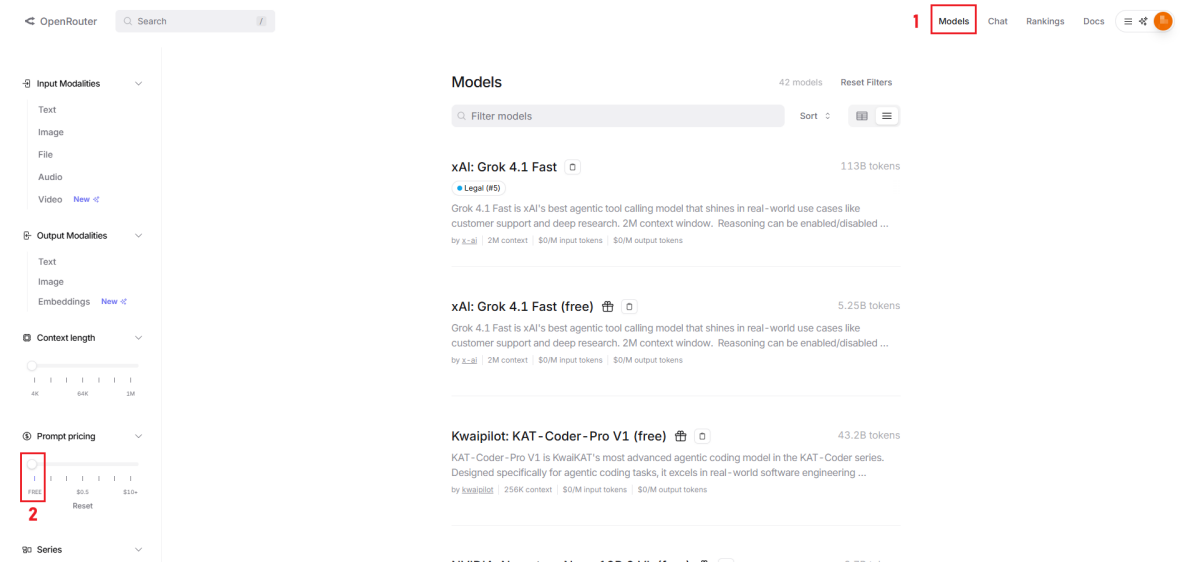
Command to close Dify:

```
sh ~/off_dify.sh
```

- For how to use these created knowledge bases, refer to the previous **[3.3.3 Configure Decision-Layer Knowledge Base]**

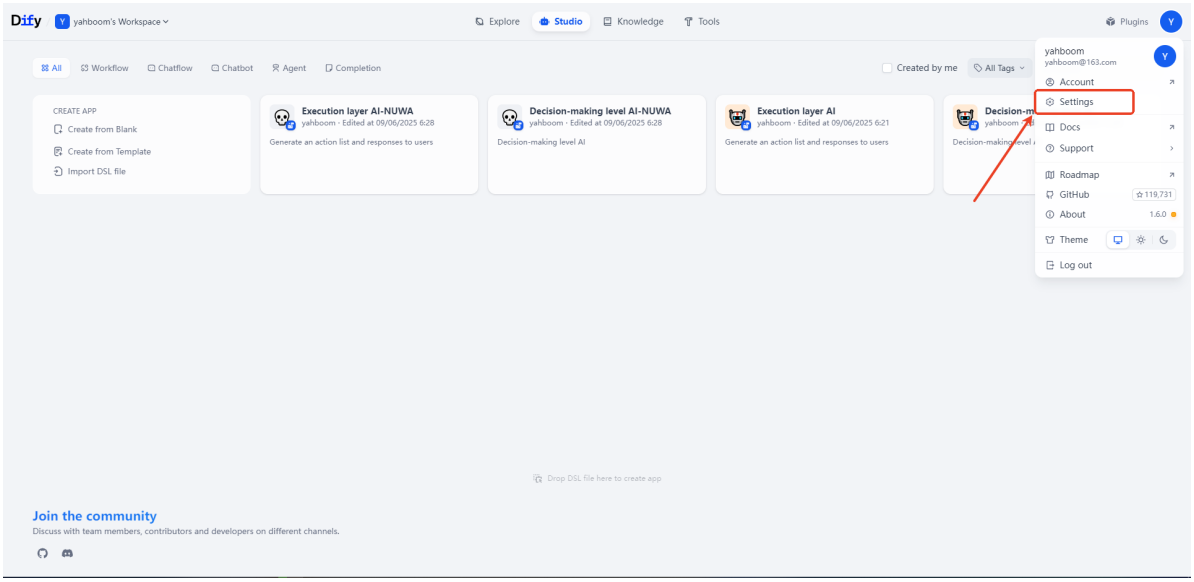
3.6 How to View Free Models on OpenRouter Platform

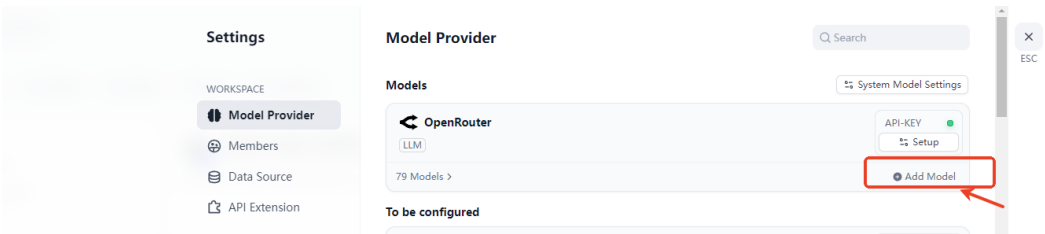
On the OpenRouter official website, click Model, select free in the Prompt pricing filter, and the free models will be displayed

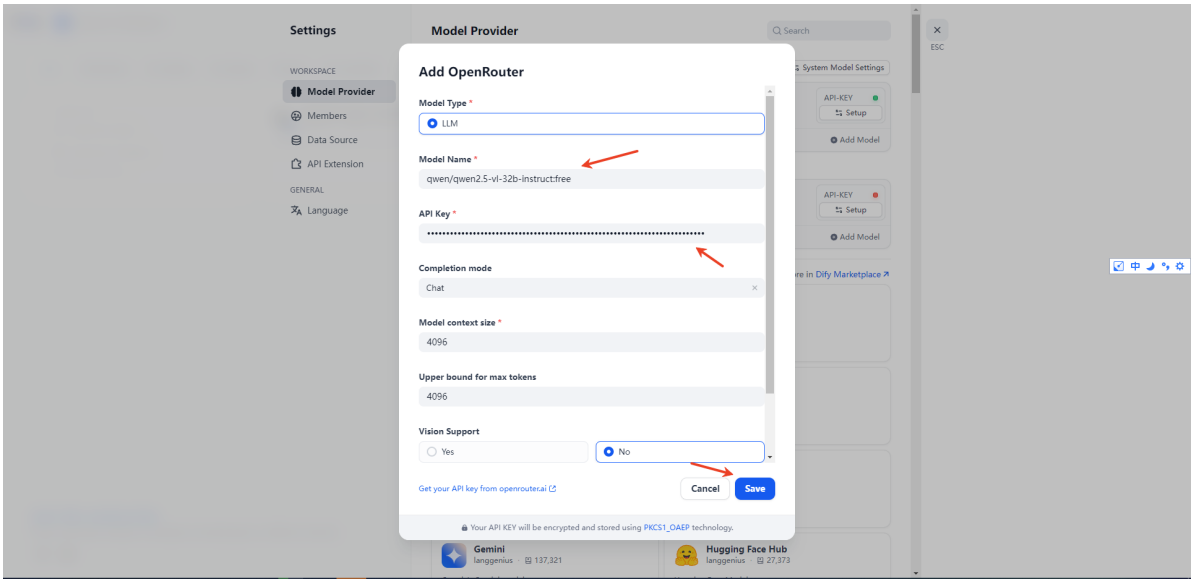


3.7 How to Import a Free Openrouter Model from the Dict.com Platform

In the Dict.com backend, locate the Openrouter large model configuration, select Import, enter the free model found in the Openrouter platform, fill in your API key, and select Import. This will allow you to use the imported model at the decision-making or execution layer.







The recommended model here is: qwen/qwen2.5-vl-32b-instruct:free. Currently, this is a relatively good large model with visual processing capabilities.

3.8 Switching Speech Models

3.8.1 Online Speech Recognition (ASR) Model and Text-to-Speech (TTS) Model

The international version uniformly uses the iFlytek Spark model provided by our company.

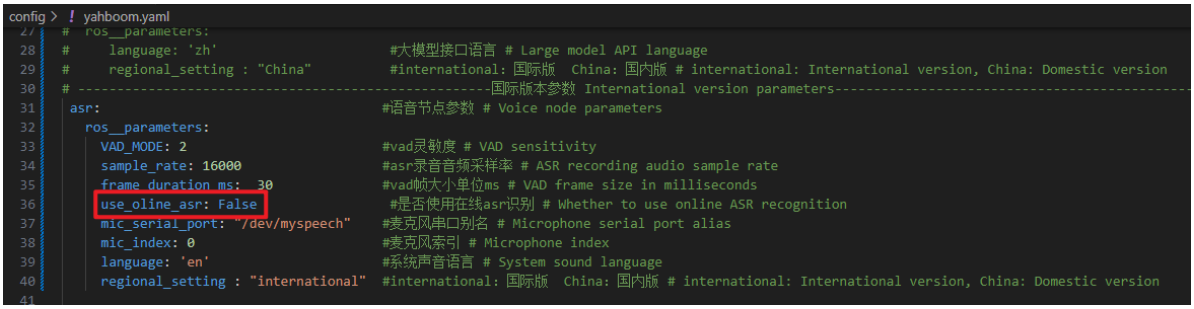
3.8.2 Local Speech Recognition Model (Orin users only)

3.8.1.1 Switching to Local Speech Recognition Model

jetson orin nano The host computer has a built-in local speech recognition model, **SenseVoiceSmall**, which can be used permanently without quota limits. You can switch to use the local speech recognition model in the yahboom.yaml configuration file. The path to the yahboom.yaml configuration file is:

```
/home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/yahboom.yaml
```

Change the **use_oline_asr** parameter in the asr node to **False**.



Then, switch to the `yahboomcar_ros2_ws/yahboomcar_ws/` workspace on the vehicle's infotainment system: Recompile the largemodel package to apply the configuration.

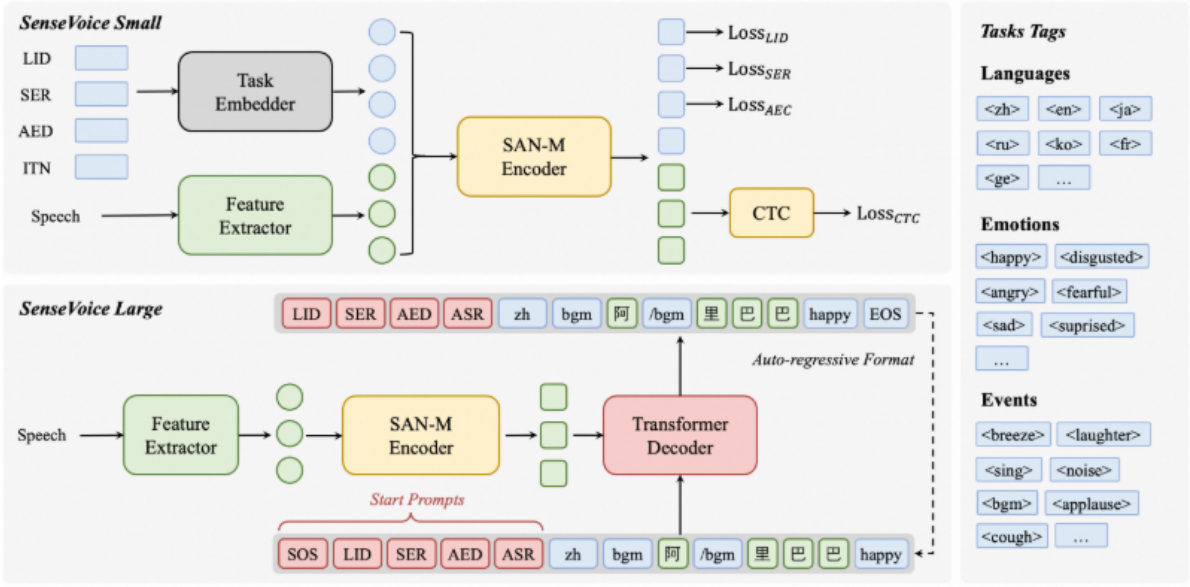
```
colcon build --packages-select largemodel
```

```
                        [System Information]
IP_Address_1: 192.168.12.66
IP_Address_2: 172.18.0.1
------------------------------------------------------
ROS_DOMAIN_ID: 61 | ROS: humble
my_robot_type: M1 | my_lidar: tmini | my_camera: nuwa
------------------------------------------------------
jetson@yahboom:~$ cd yahboomcar_ros2_ws/yahboomcar_ws/
jetson@yahboom:~/yahboomcar_ros2_ws/yahboomcar_ws$ colcon build --packages-select largemodel
Starting >>> largemodel
Finished <<< largemodel [5.37s]

Summary: 1 package finished [6.31s]
jetson@yahboom:~/yahboomcar_ros2_ws/yahboomcar_ws$
```

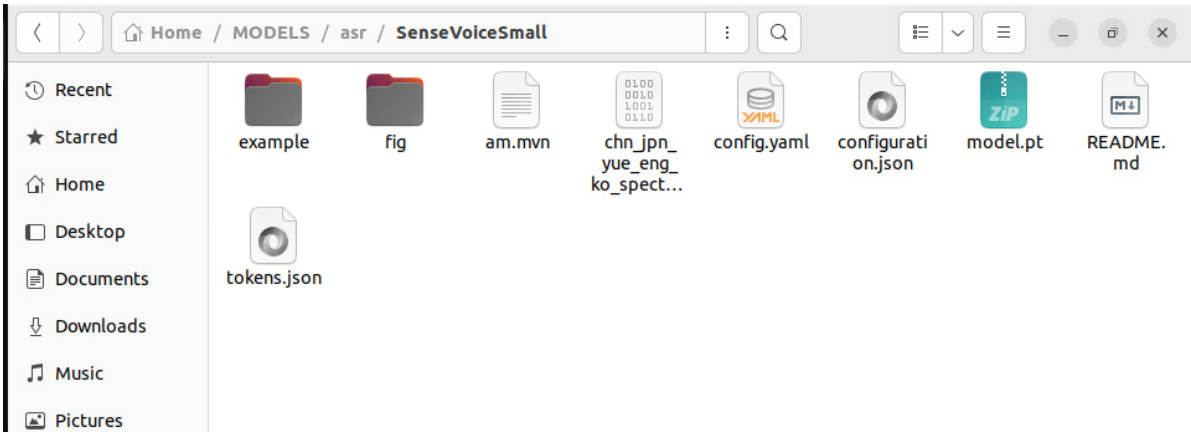
3.8.1.2 Introduction to SenseVoiceSmall

SenseVoiceSmall is an open-source model developed by Tongyi Labs. It's a multilingual audio understanding model with capabilities including speech recognition, language identification, and acoustic event detection.



SenseVoiceSmall model file location:

```
/home/jetson/MODELS/asr/SenseVoiceSmall
```



Model file address: https://www.modelscope.cn/models/iic/SenseVoiceSmall

GitHub repository: https://github.com/FunAudioLLM/SenseVoice

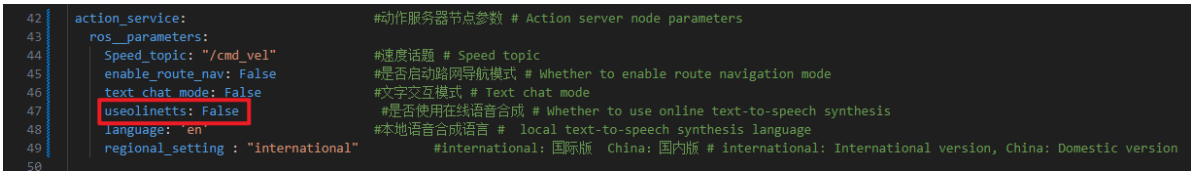
3.8.3 Local Speech Synthesis Model (Orin users only)

3.8.3.1 Switching to Local Speech Synthesis Model

jetson orin nano The host computer has a built-in local speech synthesis model, Piper, which can be used permanently without any usage limits. You can switch to use the local speech recognition model in the yahboom.yaml configuration file. The path to the yahboom.yaml configuration file is:

```
/home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/yahboom.yaml
```

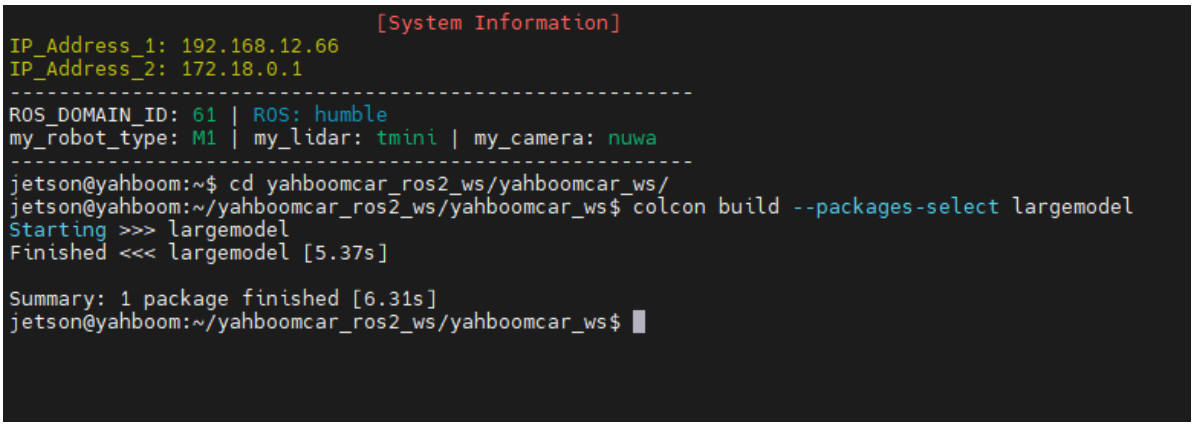
Change the useolinetts parameter of the model_service node to False



Then, switch to the `yahboomcar_ros2_ws/yahboomcar_ws/` workspace on the vehicle's infotainment system: Recompile the largemodel package to apply the configuration.

Jetson Orin Nano host computer:

```
cd /home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/
colcon build --packages-select largemodel
```



3.8.3.2 Introduction to piper

A fast, localized neural text-to-speech system.

GitHub repository address: https://gitcode.com/gh_mirrors/pi/piper

3.9 Use International Version Parameters

yahboom.yaml

International version users need to comment out the domestic version parameters

jetson orin nano host:

```
/home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/yahboom.yaml
```

jetson nano, Raspberry Pi host:

Need to enter docker first

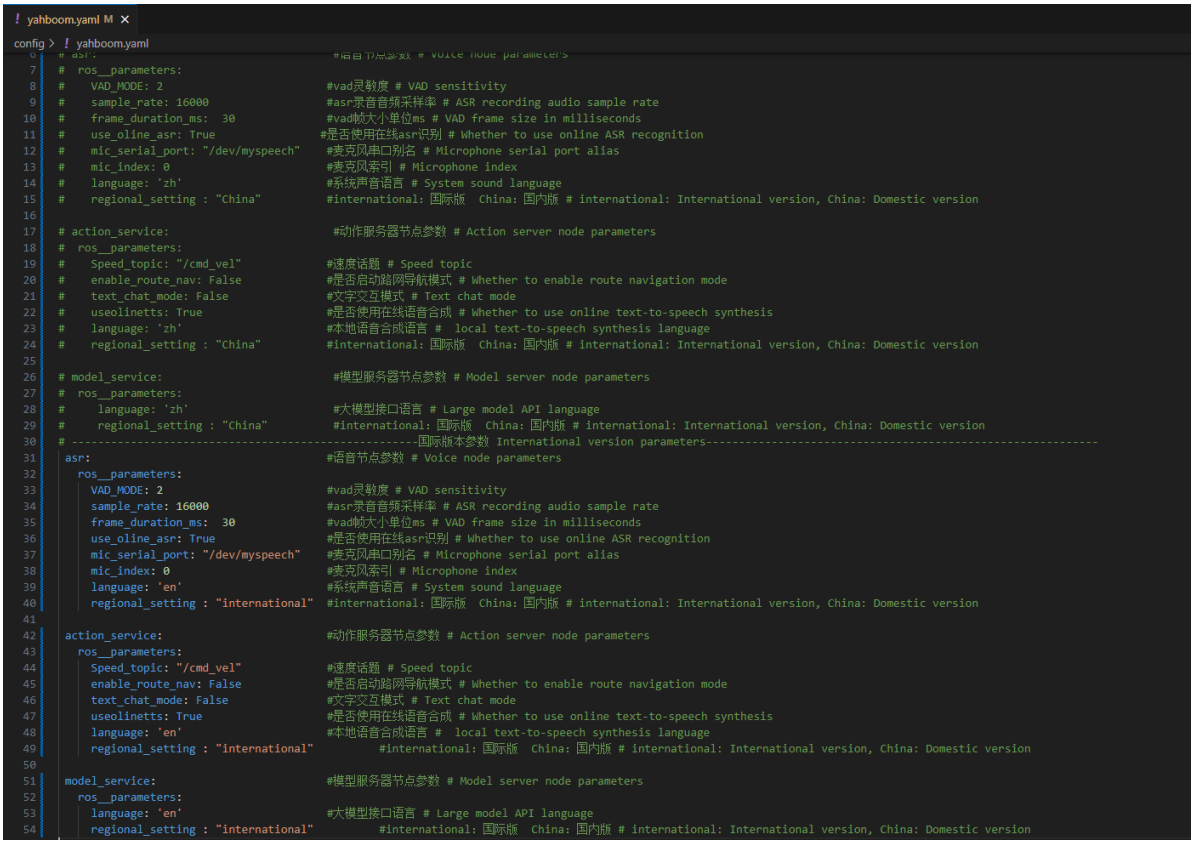
```
/root/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/yahboom.yaml
```

RDKX5 host:

```
/home/sunrise/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/yahboom.yaml
```

The commented out domestic version content is shown in the figure below:

Note: The following API configuration interfaces are all modified in vscode. If garbled characters appear when modifying in vnc, it is recommended to use the gedit editor in vnc to modify the API



large_model_interface.yaml

USB camera/depth camera users need to select the configuration corresponding to their device in the configuration file, **api-key is a fixed configuration and does not need to be replaced by users**

jetson orin nano host:

```
/home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/large_model_interface.yaml
```

RDKX5 host:

```
/home/sunrise/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/large_model_interface.yaml
```

jetson nano, Raspberry Pi host:

Need to enter docker first,

```
/root/yahboomcar_ros2_ws/yahboomcar_ws/src/largemodel/config/large_model_interface.yaml
```

```
#USB camera users
decision_AI_api_key : "app-iztrZsJiEq470poPLfnKHoRW"
execution_AI_api_key : "app-4ZA5OUWQirPc3zJHCwY6sSDe"
#Depth camera users
decision_AI_api_key : "app-1DGnalUnc4VhICrGGh0awvcH"
execution_AI_api_key : "app-wi3fvhfFGjhtIrrwSgs6RDdx"
```

USB camera user configuration is as follows, need to comment out the nuwa depth camera configuration

Depth camera user configuration is as follows, need to comment out the USB camera configuration

Then compile the function package to save the configuration:

After modifying the large model configuration parameters each time, you need to recompile the function package for the configuration to take effect. The method is as follows:

```
cd ~/yahboomcar_ros2_ws/yahboomcar_ws/
```

Switch to the `yahboomcar_ros2_ws/yahboomcar_ws/` workspace in the car terminal and recompile the largemodel function package for the configuration to take effect

```
colcon build --packages-select largemodel
```