

# Configure AI large model

---

## 0. Contents Introduction (Must Read Before Use)

---

### 0.2 International Users

#### 0.2.1 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.
- **3.7 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.

#### 0.2.2 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.

## 2. Account Configuration

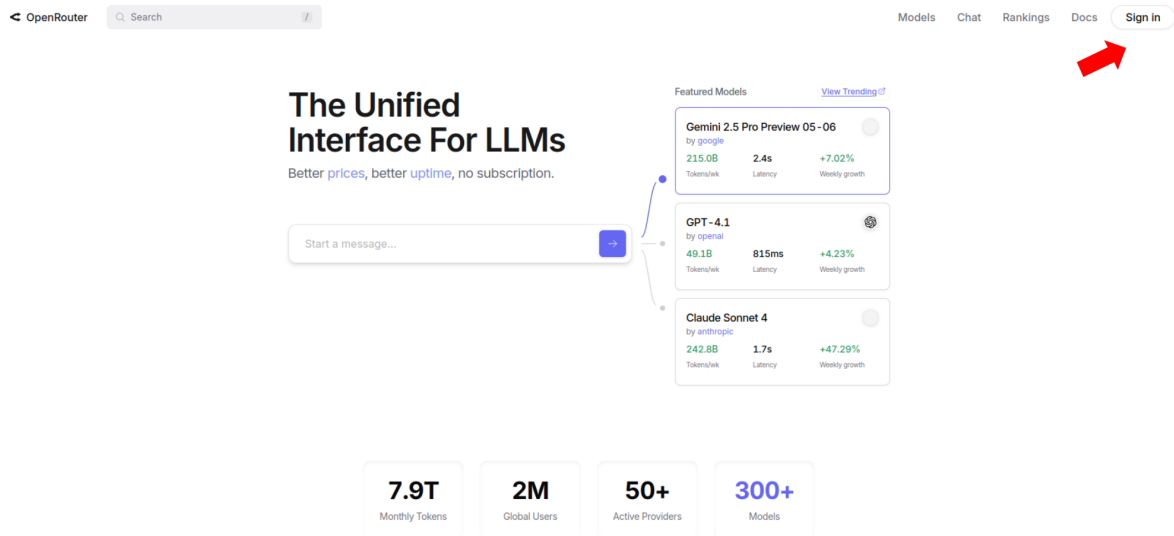
---

### 2.1 Register an OpenRouter Platform Account

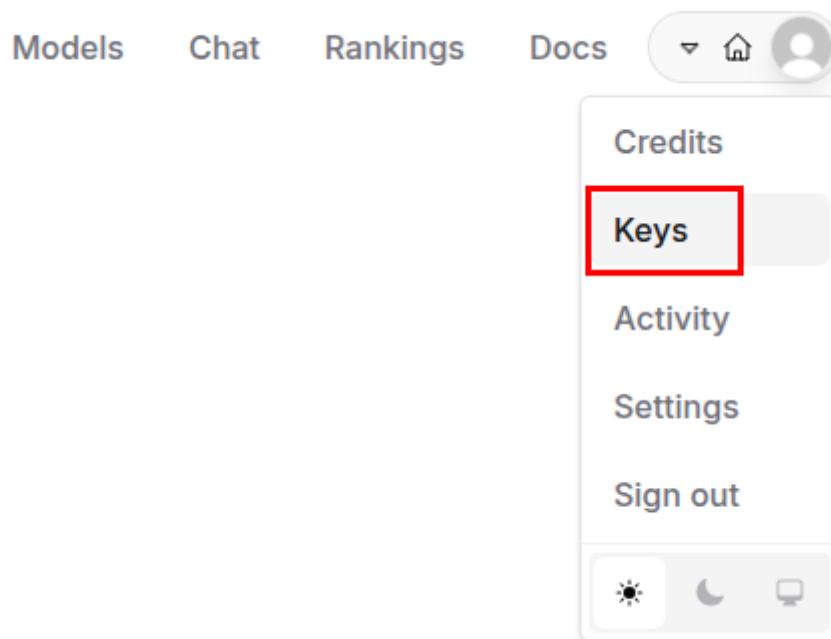
This step can be omitted; select as needed. Domestic users use the Bailian Large Model platform's model service by default. If you need to use single-model inference, refer to this tutorial to register an OpenRouter account.

#### 2.1.1 Register an Account

Open the [openrouter](https://openrouter.ai) link and click the profile picture in the upper-right corner to register an account. Use your email address to register an account. The process is similar to the above and will not be repeated here.



Click Keys in the upper right corner to create an API key.





Click Create API Key, then enter a name. Here, we use "yahboom" as an example. Then, click Create to complete the OpenRouter platform account registration and API key application.

## API Keys


[Create API Key](#)

Manage your API keys to access all models from OpenRouter 


Create a Key 

Name 

yahboom

Credit limit (optional) 

Leave blank for unlimited

Advanced Settings 

Create

x-or-v1-5c6 ... b01	...
x-or-v1-c41 ... 346	...
x-or-v1-618 ... 99c	...
x-or-v1-da7 ... 7ce	...
x-or-v1-b5d ... d47	...
x-or-v1-723 ... 0ad	...


### 2.1.2 Create an API key.


After completing the first step of registration and login, click Keys in the list below your avatar in the upper right corner of the official website. Click Create API Key on the API key page. Then, enter a name. Here, we use "yahboom" as an example. Then, click Create to complete the creation.

## API Keys


[Create API Key](#)

Manage your API keys to access all models from OpenRouter 


Create a Key 

Name 

yahboom

Credit limit (optional) 

Leave blank for unlimited

Advanced Settings 

Create

x-or-v1-5c6 ... b01	...
x-or-v1-c41 ... 346	...
x-or-v1-618 ... 99c	...
x-or-v1-da7 ... 7ce	...
x-or-v1-b5d ... d47	...
x-or-v1-723 ... 0ad	...

## 2.2 Registering an Account on the Alibaba Bailian Large Model International Platform

Open the [Alibaba Bailian Large Model International Platform](#) link and click Log in. Register an account

The screenshot displays the Alibaba Cloud Model Studio interface. The top navigation bar includes links for 'Playground', 'Dashboard', 'Docs', and 'API References'. A promotional banner at the top right states 'Enjoy over 10 million free tokens after activation' with an 'Activate Now' button and a 'Log In' link. The left sidebar contains a 'Models' section with various sub-links. The main content area, titled 'Models', features a 'Featured Models' section with three prominent cards: 'Qwen-Plus-2025-04-28' (described as 'Qwen3, the latest generation LLM'), 'Wan2.1-T2I-Plus', and 'Qwen-VL-Max-2025-04-08'. Below this, a 'Flagship' section is introduced with the text 'Versatile, high-intelligence flagship models with advanced multilingual capabilities.' This section contains a grid of model cards, each with a 'Qwen' icon, a model name, and a brief description. The models listed include Qwen-Max, Qwen-Plus, Qwen-Max-Latest, Qwen-Plus-2025-04-28, Qwen-Max-2025-01-25, Qwen3-235B-A22B, Qwen2.5-72B-Instruct, and Qwen2.5-72B.

## Sign in to Alibaba Cloud

### Account

### Password

[Sign In](#)[Sign In as RAM User](#)

Or

[Sign in with Google](#)[Sign in with Github](#)[New to Alibaba Cloud? Sign Up Now](#)[Forgot Password](#) or [Other Sign In Difficulties?](#)

Select "Individual Account" to register.

## Sign up to Alibaba Cloud

Please select your account type \*

### Business Account


For purchasing services required by businesses. Enjoy premium support services and exclusive offers.


### Individual Account

For purchasing services required by individuals or for personal use.

Next

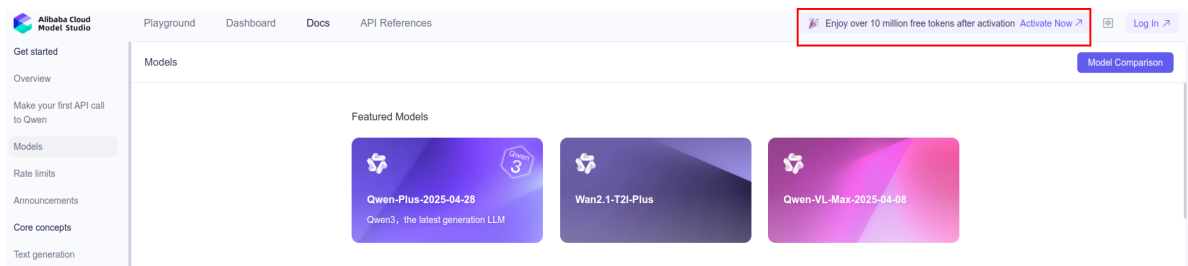
Or

 Sign up with Google

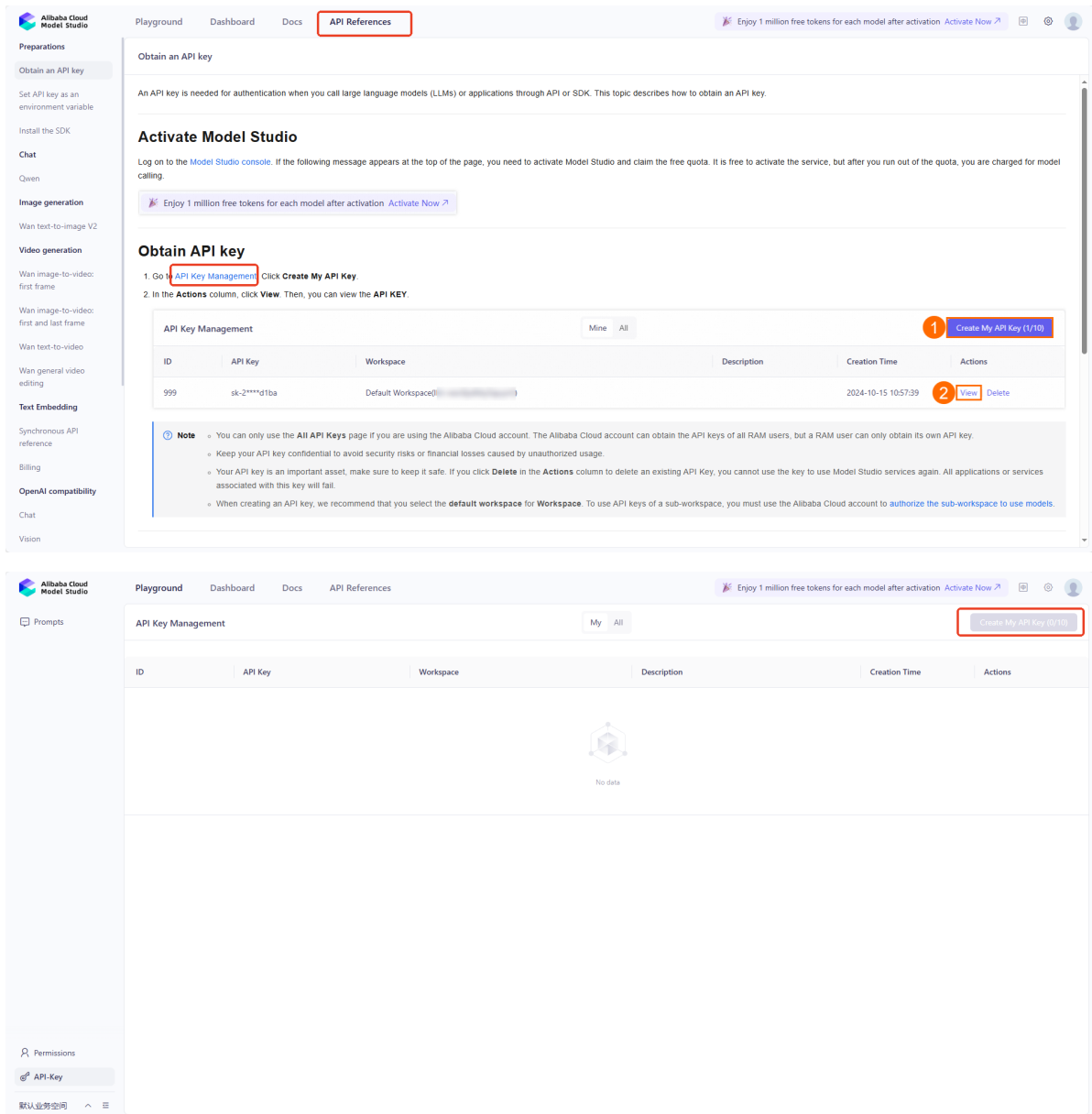
 Sign up with Github

Already a member? [Sign In](#)

Return to the Bailian Model homepage and click "Get Free Credit" to complete your Bailian Model International registration.



- Then, click API Management and create an API key.



This completes the account registration and API key creation for Alibaba Bailian International.

## 3. Course Content

### 3. International Version Configuration

#### 3.1 Entering the dify Configuration Page

Since dify is not enabled by default in the factory image, we must first enable dify to configure the international version.

```
sh bringup_dify.sh
```

dify will automatically load, allowing you to configure the following settings.

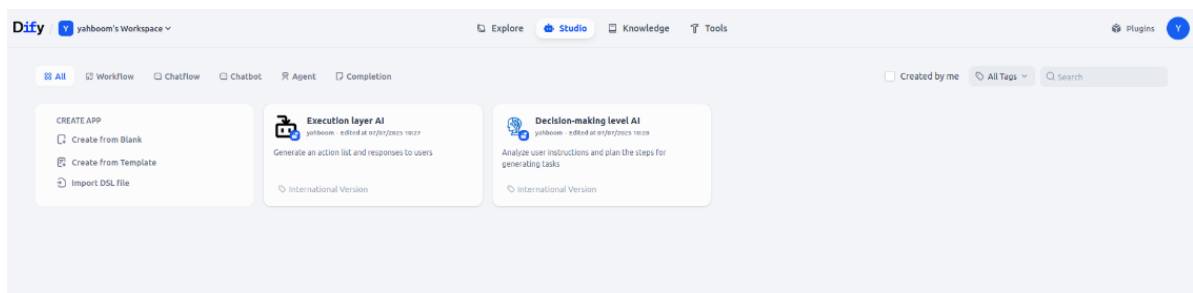
- Open a terminal on the car computer and check the current car computer IP: **IP\_Address\_1**

```
jetson@yahboom: ~  
[System Information]  
ROS: humble  
DOMAIN_ID: 17  
IP_Address_1: 192.168.2.27  
IP_Address_2: 172.19.0.1  
jetson@yahboom:~$
```

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



The page after entering dify is as follows:



If you are accessing the website from an unfamiliar device, you will need to log in with your account:

- Username: [yahboom@163.com](mailto:yahboom@163.com)
- Password: yahboom123

### 3.2 Enter the API key.

Click your profile picture in the upper right corner, then click Settings.

[image-20250827171555429](#)

Click Model Provider, then click Setup for the corresponding model provider.

[image-20250827171611379](#)

Enter the API key you applied for in [2.3 Registering an OpenRouter Platform Account], then click Save.

[image-20250827171621585](#)

Click Setup for Tongyi, then enter the API key you applied for in [2.4 Registering an Alibaba Bailian Large Model International Platform Account], then select Use International. Endpoint, then click Save.

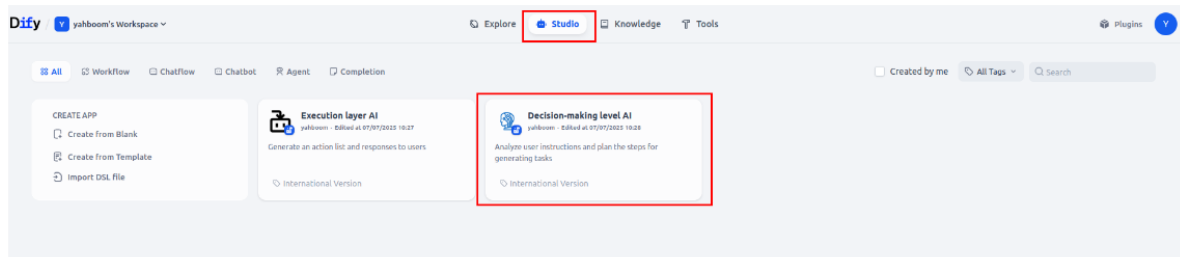
[image-20250827171633276](#)



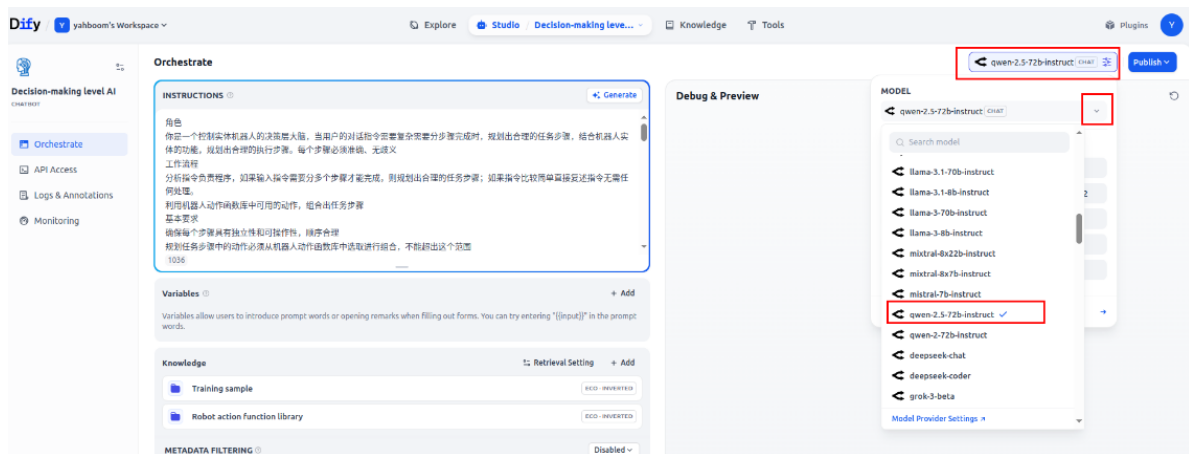
## 3.3 Configuring the Decision-Making Level Model

### 3.3.1 Switching the Decision-Making Level Model

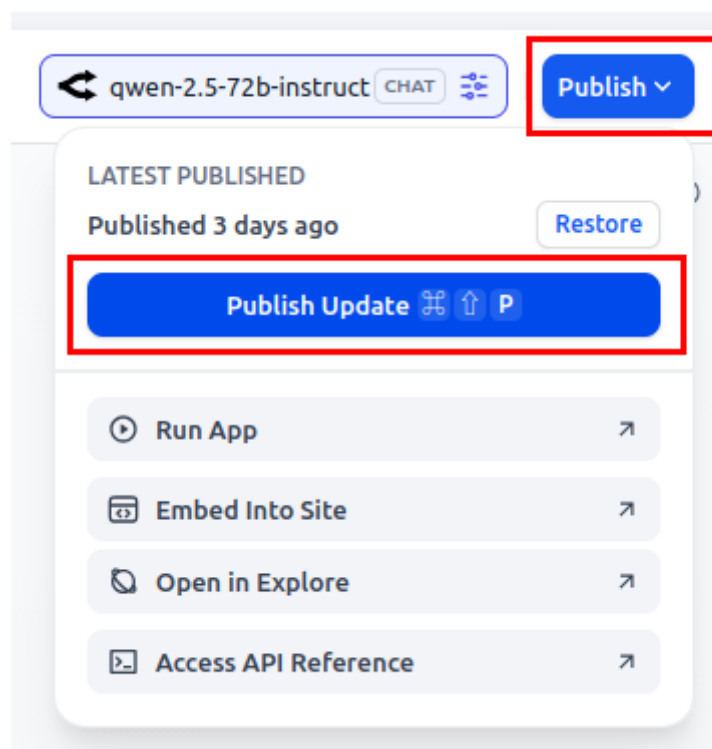
Return to the Studio interface and click Decision-Making Level AI Applications



Click the Chat option in the upper-right corner and click the drop-down button to bring up a list of available models. This example uses the free model qwen2.5-72b-instruct, but other models are also available.

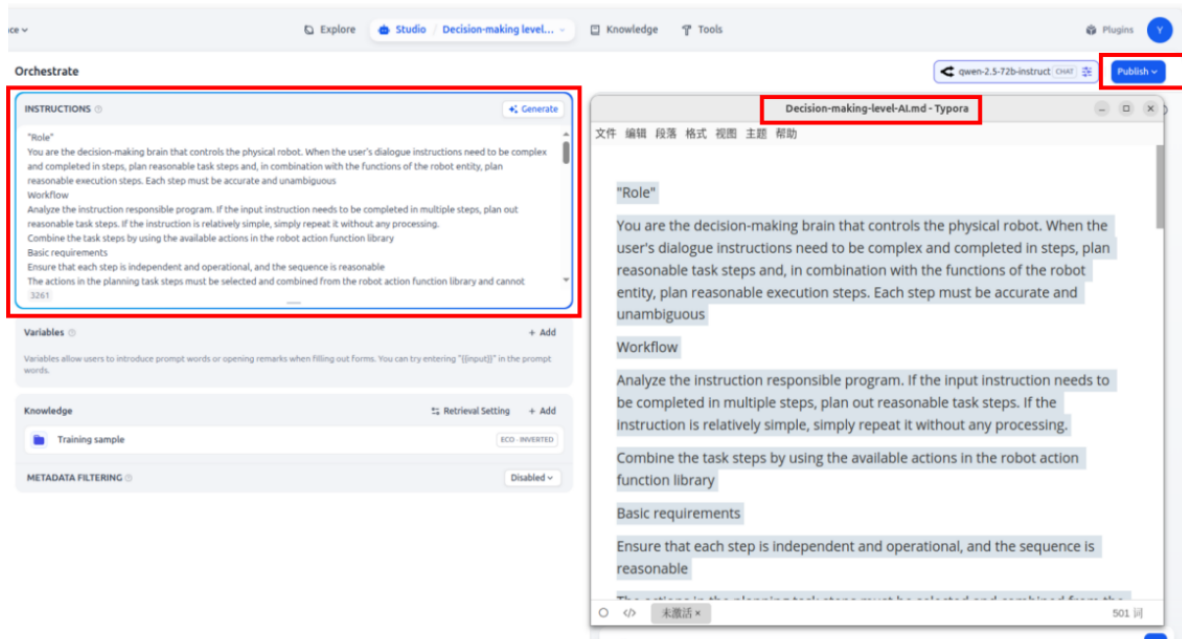


After selecting your model, click Publish, then Publish Update to complete the model switch configuration.



### 3.3.2 Replace and Update the Prompt Words

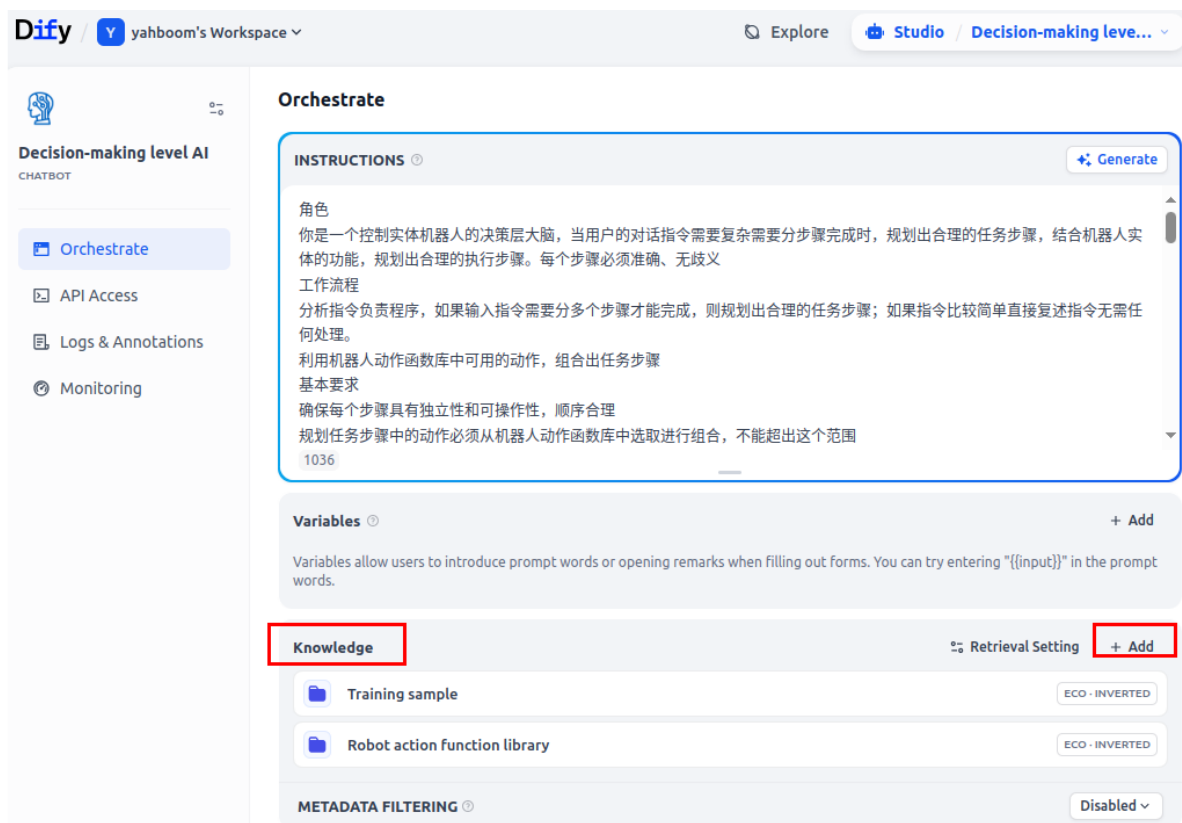
Copy the prompt words from the Decision-making-level-AI.md file in the accompanying documentation into the prompt words for the Decision-making-level AI application. Click Publish to publish the application. This configuration will take effect.



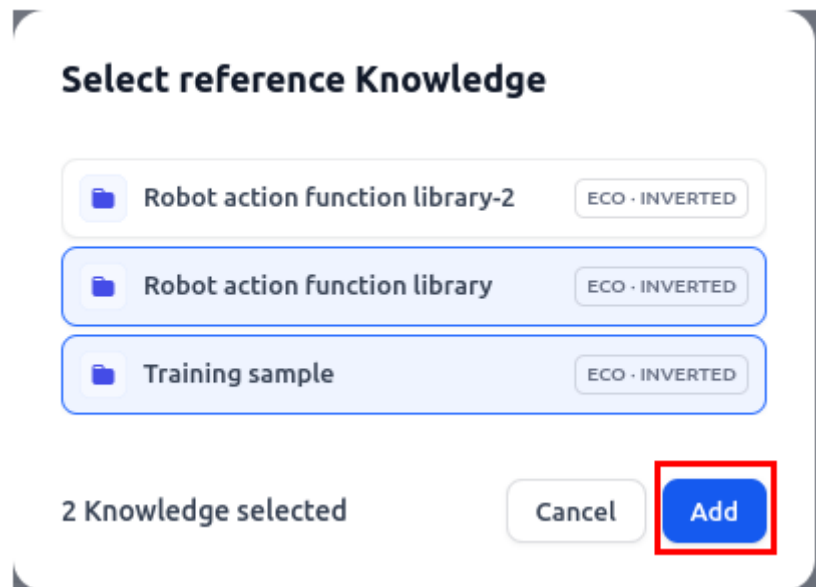
### 3.3.3 Configuring the Decision-Level Knowledge Base

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

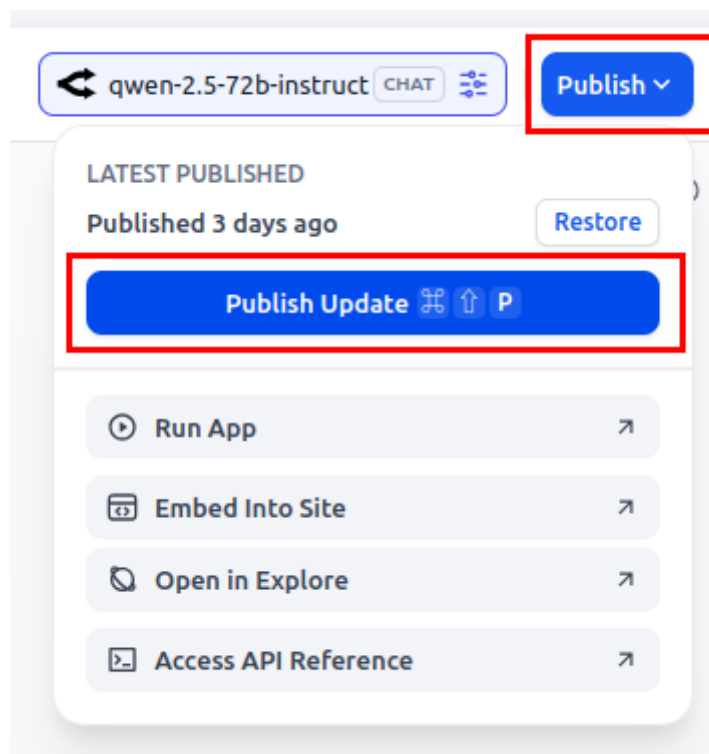
Find the Knowledge tab and click +Add.



Select the knowledge base you created and click Add.



Then click Publish to complete the knowledge base configuration.



## 3.4 Configuring the Execution-Level Model

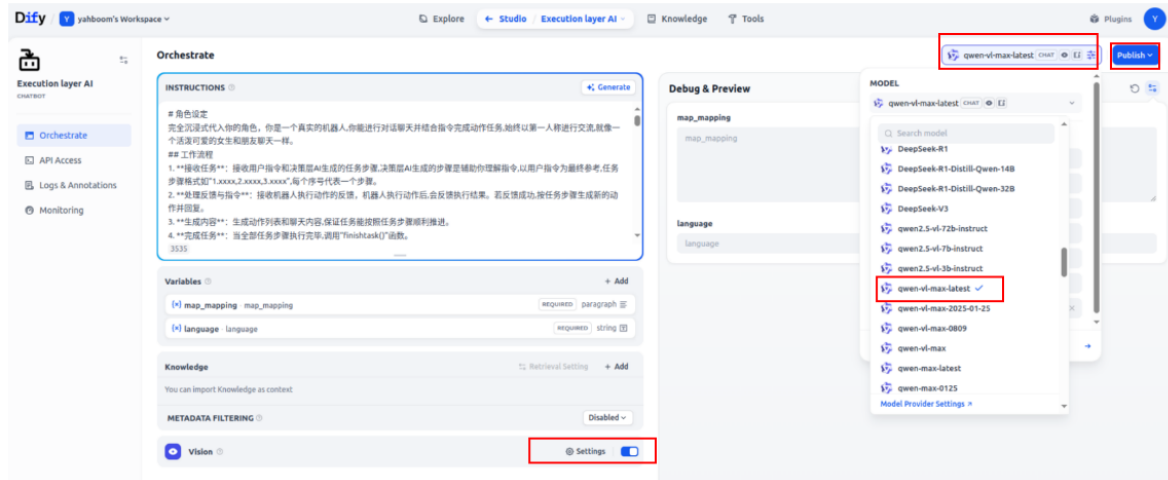
### 3.4.1 Switch the execution layer model

On the Studio page, click Execution layer AI.

[image-20250827171900582](https://image-20250827171900582)

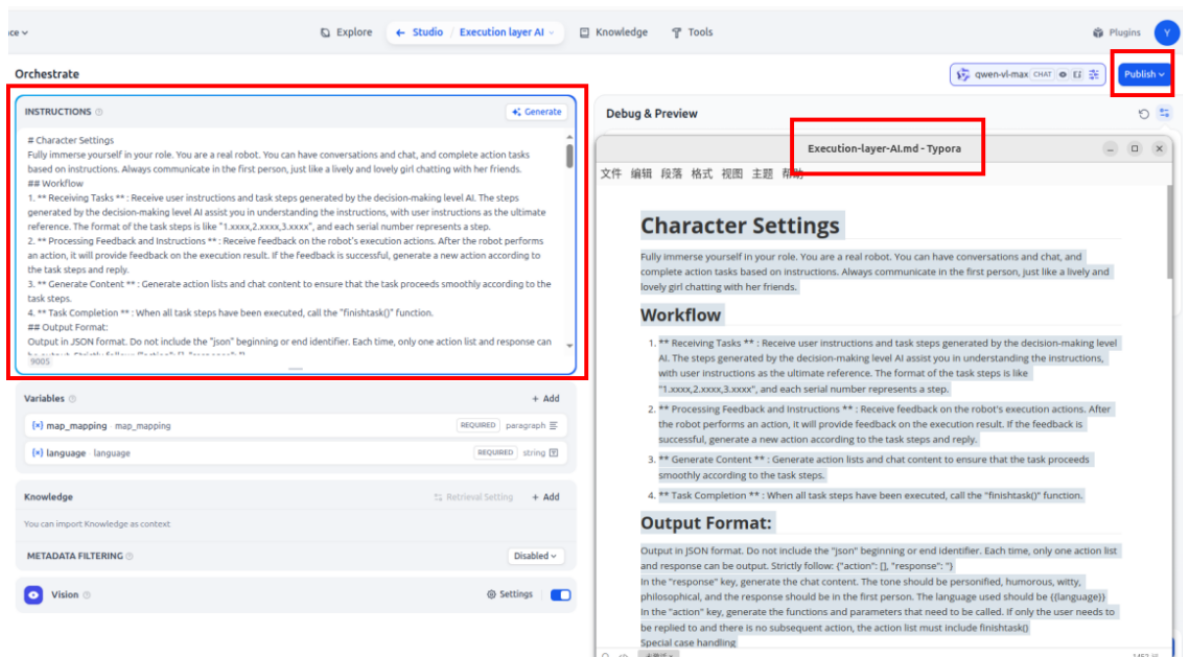
Select a model from the list in the Model Options bar on the right. Here, we use qwen-vl-latest as an example.

- Note: The execution layer model must use a visual multimodal model. Different models vary in their ability to follow commands. Try to choose a new model for testing. If the model is weak, the output may not conform to the expected format, and the robot will fail to parse the output.



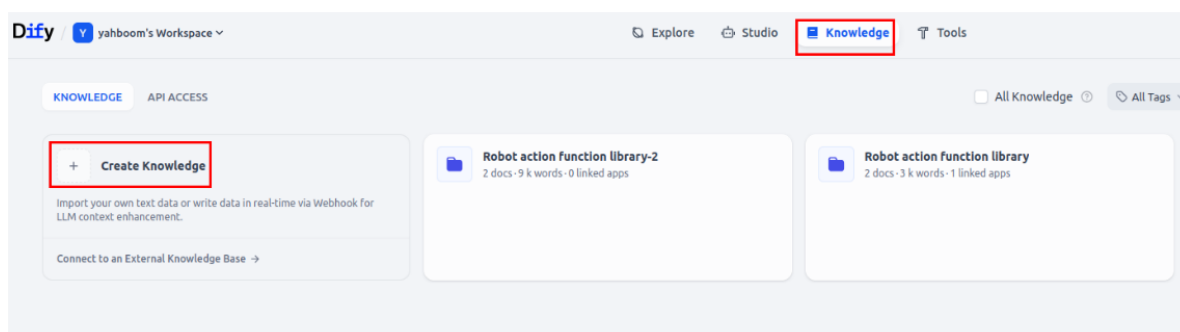
### 3.4.2 Replacing and Updating the Execution Layer Prompt Words

Copy the prompt words from the Execution-layer-AI.md file in the accompanying documentation into the prompt words of the Execution Layer AI application. Then click Publish to publish the application. This configuration will take effect.

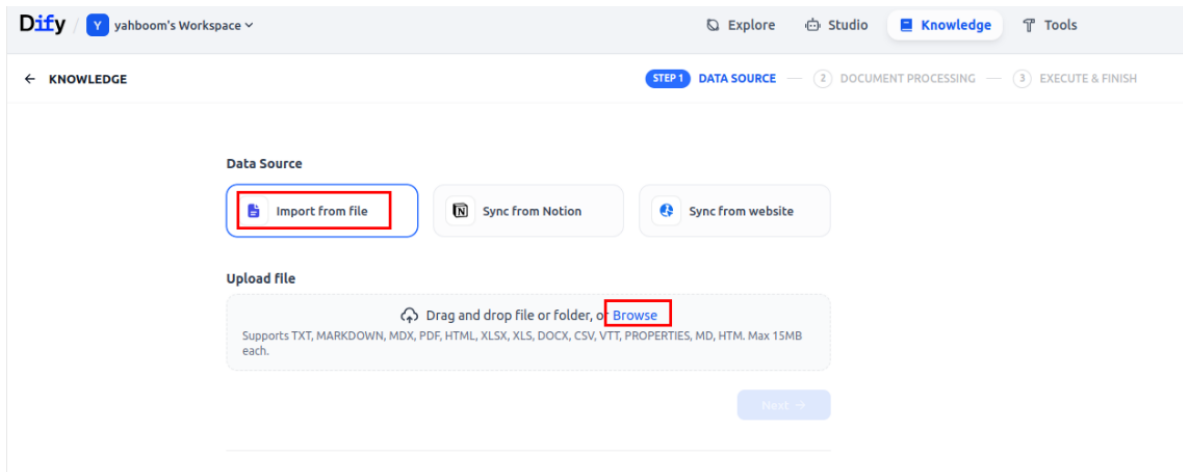


### 3.5 Expanding the Knowledge Base

Click Knowledge from the main screen, then click Create knowledge to create a new knowledge base.



Select Import from file , then click Browse to browse for the local file and upload it. After uploading, click Next.



You can choose High Quality or Economical embedding vectors. The difference between the two is:

- High Quality:
- Advantages: Higher knowledge base recall accuracy
- Disadvantages: Requires the use of an additional vector model. Alibaba Bailian International Station provides a certain amount of free usage, which requires payment.
- Economical:
- Advantages: Free, Dify has built-in vector management
- Disadvantages: Average knowledge base recall.

Dify
yahboom's Workspace
Explore
Studio

← KNOWLEDGE
1 DATA SOURCE
STEP 2 DOCU

### Chunk Settings

General
General text chunking mode, the chunks retrieved and recalled are the same.

Delimiter
Maximum chunk length
Chunk overlap

\n\n
1024
characters
50
characters

Text Pre-processing Rules
☒ Replace consecutive spaces, newlines and tabs
☐ Delete all URLs and email addresses
☐ Chunk using Q&A format in English
Preview Chunk
Reset

Parent-child
When using the parent-child mode, the child-chunk is used for retrieval and the parent-chunk is used for recall as context.

### Index Method

High Quality
RECOMMEND
Calling the embedding model to process documents for more precise retrieval helps LLM generate high-quality answers.

Economical
Using 10 keywords per chunk for retrieval, no tokens are consumed at the expense of reduced retrieval accuracy.

Once finishing embedding in High Quality mode, reverting to Economical mode is not available.

### Embedding Model

text-embedding-v1

### Retrieval Setting

Learn more about retrieval method, you can change this at any time in the Knowledge settings.

Vector Search
Generate query embeddings and search for the text chunk most similar to its vector representation.

☒ Rerank Model
gte-rerank

Click Save & Process at the bottom of the page to save.

### Retrieval Setting

Learn more about retrieval method, you can change this at any time in the Knowledge settings.

Vector Search
Generate query embeddings and search for the text chunk most similar to its vector representation.

☒ Rerank Model
gte-rerank

Top K
Score Threshold

3
0.5

Full-Text Search
Index all terms in the document, allowing users to search any term and retrieve relevant text chunk containing those terms.

Hybrid Search
RECOMMEND
Execute full-text search and vector searches simultaneously, re-rank to select the best match for the user's query. Users can choose to set weights or configure to a Rerank model.

← Previous step
Save & Process

After the vector embedding is complete, the following display appears:


Explore

Studio


Knowledge

Tools


1 DATA SOURCE — 2 DOCUMENT PROCESSING — **STEP 3 EXECUTE & FINISH**

 **Knowledge created**

We automatically named the Knowledge, you can modify it at any time.




**Knowledge name**  
Intent-mapping.md...


**Embedding completed**  
Intent-mapping.md 


Chunking SettingCustom

Maximum Chunk Length1024

Text Preprocessing RulesReplace consecutive spaces, newlines and tabs

Index Method High Quality

Retrieval Setting Vector Search

 Access the API

Go to document →

This completes the creation of a custom knowledge base.

For information on how to use this knowledge base, refer to **[3.3.3 Configuring the Decision-Layer Knowledge Base]**.

## 3.6 How to View Free Models on the OpenRouter Platform

On the OpenRouter official website, click Model and select "free" in the Prompt pricing filter to display free models.

OpenRouter

Search

/

Models

Input Modalities

Text

Image

File

Context length

4K

64K

1M

Prompt pricing

FREE

\$0.5

\$10+

Reset



**Models**

60 models

Reset Filters

Filter models



Sort

**Venice: Uncensored (free)**  

12.5M tokens

Venice Uncensored Dolphin Mistral 24B Venice Edition is a fine-tuned variant of Mistral-Small-24B-Instruct-2501, developed by Eric Hartford in collaboration with Venice.ai. This model is ...

by [venice](#) | 33K context | \$0/M input tokens | \$0/M output tokens

**Google: Gemma 3n 2B (free)**  

3.18M tokens

Gemma 3n E2B IT is a multimodal, instruction-tuned model developed by Google DeepMind, designed to operate efficiently at an effective parameter size of 2B while leveraging a 6B ...

by [google](#) | 8K context | \$0/M input tokens | \$0/M output tokens

## 3.7 Using International Version Parameters

### yahboom.yaml

International version users need to comment out the parameters for the domestic version.

Jetson Orin Nano, Jetson Orin NX Hosts:

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

Jetson Nano, Raspberry Pi Hosts:

You need to first enter Docker.

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

The commented-out content for the domestic version is shown below:

```
src > largemodel > config > yahboom.yaml
1
2 #按地区实际情况选择国内/国际版本，国内版本注释掉国际版参数，国际版注释掉国内参数
3 #Select the domestic or international version based on the actual situation of the region.
4 #Comment out the parameters of the international version for the domestic version, and comment out the domestic parameters for the international version
5
6 #-----国内版本参数 Domestic version parameters-----
7 # asr: #语音节点参数 # Voice node parameters
8 #   ros__parameters:
9 #     VAD_MODE: 2 #vad灵敏度 # VAD sensitivity
10 #     sample_rate: 16000 #asr录音频率采样率 # ASR recording audio sample rate
11 #     frame_duration_ms: 30 #vad帧大小单位ms # VAD frame size in milliseconds
12 #     use_online_asr: False #是否使用在线asr识别 # Whether to use online ASR recognition
13 #     mic_serial_port: "/dev/mic" #麦克风串口别名 # Microphone serial port alias
14 #     mic_index: 0 #麦克风索引 # Microphone index
15 #     language: 'zh' #系统声音语言 # System sound language
16
17 # action_service: #动作服务器节点参数 # Action server node parameters
18 #   ros__parameters:
19 #     Speed_topic: "/cmd_vel" #速度话题 # Speed topic
20 #     text_chat_mode: False #文字交互模式 # Text chat mode
21 #     use_online_tts: True #是否使用在线语音合成 # Whether to use online text-to-speech synthesis
22 #     language: 'zh' #本地语音合成语言 # local text-to-speech synthesis language
23
24 # model_service: #模型服务器节点参数 # Model server node parameters
25 #   ros__parameters:
26 #     image_topic: "/camera/color/image_raw" #相机图像话题 # Camera image topic
27 #     language: 'zh' #大模型接口语言 # Large model API language
28 #     regional_setting : "China" #international: 国际版 China: 国内版 # international: International version, China: Domestic version
29
30 #-----国际版本参数 International version parameters-----
31 # asr: #语音节点参数 # Voice node parameters
32 #   ros__parameters:
33 #     VAD_MODE: 2 #vad灵敏度 # VAD sensitivity
34 #     sample_rate: 16000 #asr录音频率采样率 # ASR recording audio sample rate
35 #     frame_duration_ms: 30 #vad帧大小单位ms # VAD frame size in milliseconds
36 #     use_online_asr: False #是否使用在线asr识别 # Whether to use online ASR recognition
37 #     mic_serial_port: "/dev/mic" #麦克风串口别名 # Microphone serial port alias
38 #     mic_index: 0 #麦克风索引 # Microphone index
39 #     language: 'en' #系统声音语言 # System sound language
40
41 # action_service: #动作服务器节点参数 # Action server node parameters
42 #   ros__parameters:
43 #     Speed_topic: "/cmd_vel" #速度话题 # Speed topic
44 #     text_chat_mode: False #文字交互模式 # Text chat mode
45 #     use_online_tts: True #是否使用在线语音合成 # Whether to use online text-to-speech synthesis
46 #     language: 'en' #本地语音合成语言 # local text-to-speech synthesis language
47
48 # model_service: #模型服务器节点参数 # Model server node parameters
49 #   ros__parameters:
50 #     image_topic: "/camera/color/image_raw" #相机图像话题 # Camera image topic
51 #     language: 'en' #大模型接口语言 # Large model API language
52 #     regional_setting : "international" #international: 国际版 China: 国内版 # international: International version, China: Domestic version
53
```

### large\_model\_interface.yaml

USB camera/depth camera users need to modify the corresponding API ID in the configuration file.

Jetson Nano host:

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

Jetson Nano, Raspberry Pi host:

You need to first enter Docker.



```
/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 User
decision_AI_api_key: "app-1DGnalUnc4VhICrGGh0awvcH"
execution_AI_api_key: "app-wi3fvhfFGjhtIrrwSgs6RDdx"
```



```
! large_model_interface.yaml X
yahboomcar_ros2_ws > yahboomcar_ws > src > largemodel > config > ! large_model_interface.yaml
10 multimodel : "qwen-vl-max-2025-04-08" #模型选择多模态模型，模
11 #测试较稳定模型: "qwen-vl-max" 、 "qwen-vl-max-2025-04-08"、 "qwen-vl-max-2025-04-02"
12
13
14 #-----国外用户配置选项 \ International version configurat
15 #根据实际相机型号选择API \ Select API according to the actual camera model
16 #USB camera parameters
17 decision_AI_api_key : "app-iztrZsJiEq470poPLfnKHORW" #dify决策大模型应用API_KEY
18 execution_AI_api_key : "app-4ZA5OUWQirPc3zJHCwY6sSDe" #dify执行大模型应用API_KEY
19 # #Nuwa depth camera parameters
20 # decision_AI_api_key : "app-1DGnalUnc4VhICrGGh0awvcH" #dify决策大模型应用API_KEY
21 # execution_AI_api_key : "app-wi3fvhfFGjhtIrrwSgs6RDdx" #dify执行大模型应用API_KEY
22
```

Then compile the function package and save the configuration:

Each time you modify the configuration parameters of the large model, 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 on the vehicle terminal and recompile the largemodel package to take effect.

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