

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.

Configure AI large model

0. Contents Introduction (Must Read Before Use)

 0.1 International Users

 0.1.1 Required Configuration Items

 0.1.2 Optional Configuration Items

 1. Course Content

 2. Account Configuration

 2.1 Register OpenRouter Platform Account

 2.1.1 Register Account

 2.1.2 Create API-KEY

 2.2 Register Alibaba Cloud Model Studio Platform Account

 2.2.1 Register Account

 2.2.2 Free Quota Description

 2.2.3 Create API-KEY

 3. International Version Usage Configuration

 3.1 Access Dify Configuration Page

 3.2 Fill in API-KEY

 3.3 Configure Decision-Layer Model

 3.3.1 Switch Decision-Layer Model

- 3.3.2 Replace and Update Instructions
- 3.3.3 Configure Decision-Layer Knowledge Base
- 3.4 Configure Execution-Layer Model
 - 3.4.1 Switch Execution-Layer Model
 - 3.4.2 Replace and Update Execution-Layer Instructions
- 3.5 Extended Knowledge Base
- 3.6 How to View Free Models on OpenRouter Platform
- 3.7 How to Import a Free Openrouter Model from the Dict.com Platform
- 3.8 Switching Speech Models
 - 3.8.1 Online Speech Recognition (ASR) Model and Text-to-Speech (TTS) Model
 - 3.8.2 Local Speech Recognition Model (Orin users only)
 - 3.8.1.1 Switching to Local Speech Recognition Model
 - 3.8.1.2 Introduction to SenseVoiceSmall
 - 3.8.3 Local Speech Synthesis Model (Orin users only)
 - 3.8.3.1 Switching to Local Speech Synthesis Model
 - 3.8.3.2 Introduction to piper
- 3.9 Use International Version Parameters
 - yahboom.yaml
 - large_model_interface.yaml

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 `largetmodel` 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/largetmodel
```

Jetson Nano/Raspberry Pi host:

Requires entering Docker first

```
root/yahboomcar_ros2_ws/yahboomcar_ws/src/largetmodel
```

RDK X5 host:

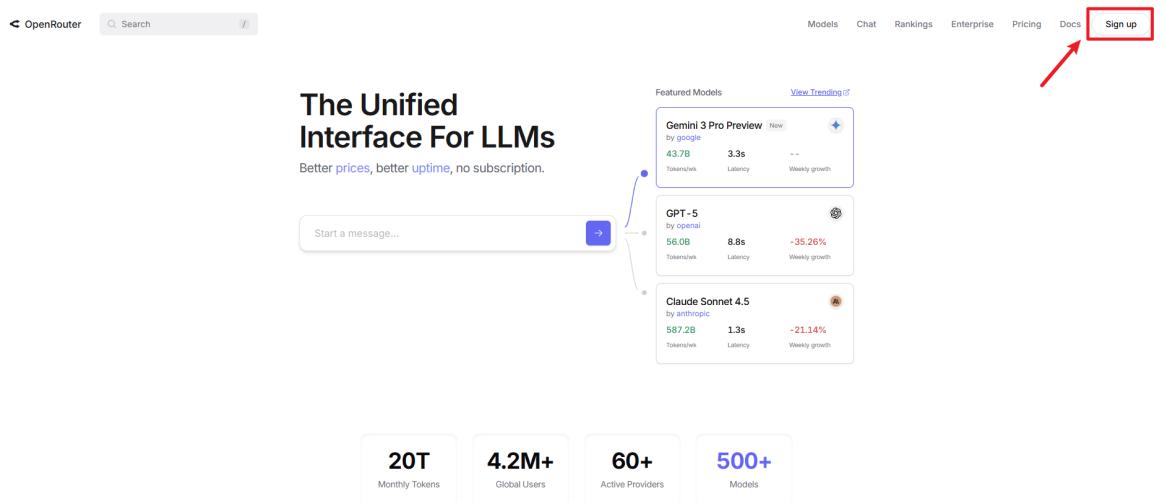
```
/home/sunrise/yahboomcar_ros2_ws/yahboomcar_ws/src/largetmodel
```

2. Account Configuration

2.1 Register OpenRouter Platform Account

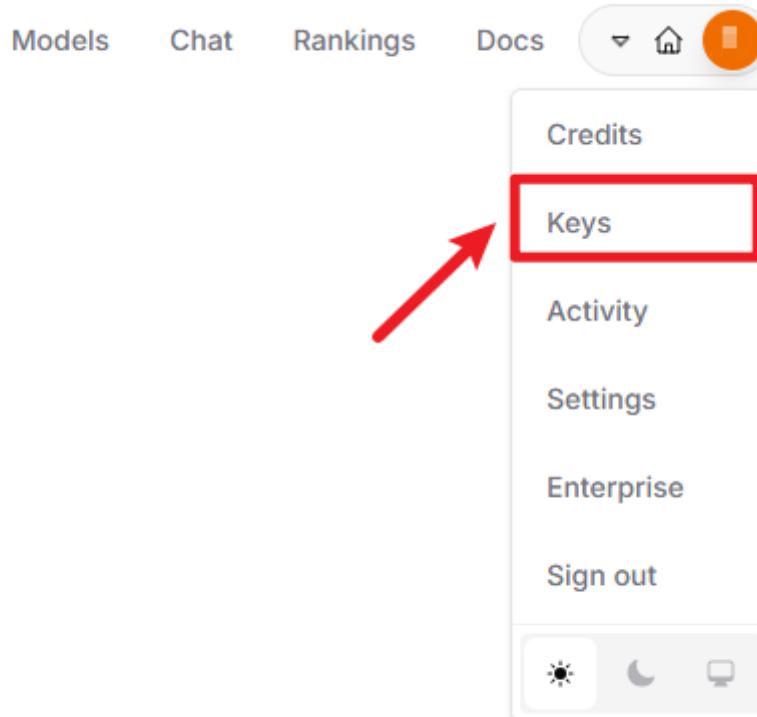
2.1.1 Register Account

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.

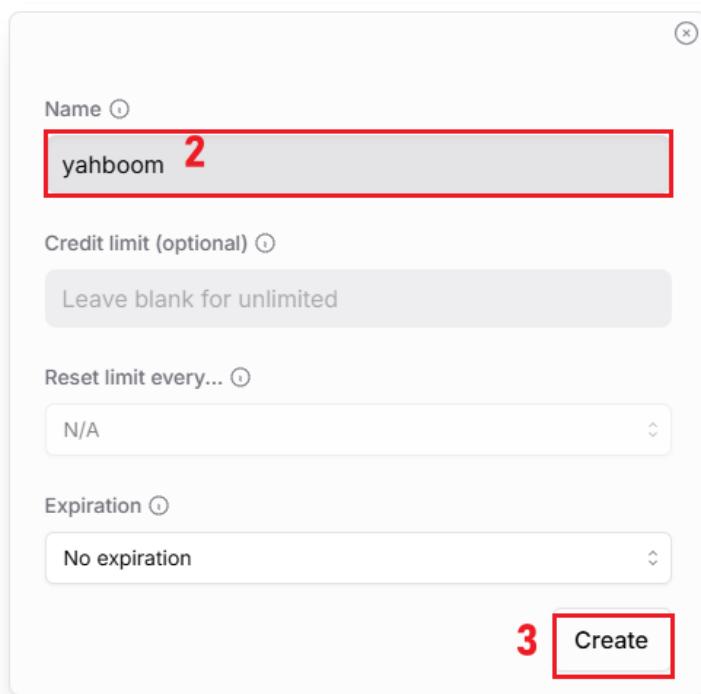


After jumping to the API-Keys page, click Create API Key, then enter any name. Here we use "yahboom" as an example, then click Create to complete the creation.

API Keys

1

Create API Key



A modal dialog box titled 'Create API Key' with a red border. It contains the following fields:

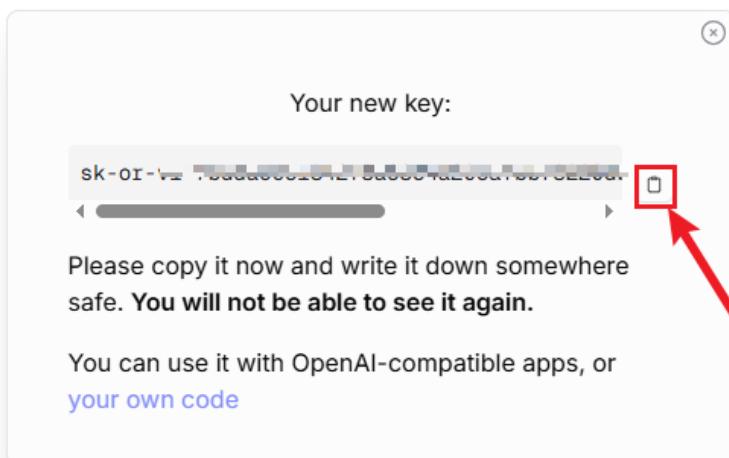
- Name: (2)
- Credit limit (optional):
- Reset limit every...:
- Expiration:
- Create** button (3)

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.**

API Keys

Create API Key

Create a new API key to access all models from OpenRouter ⓘ



A modal dialog box titled 'Your new key:' containing the following content:

Your new key:
sk-or-[REDACTED]
Please copy it now and write it down somewhere safe. **You will not be able to see it again.**

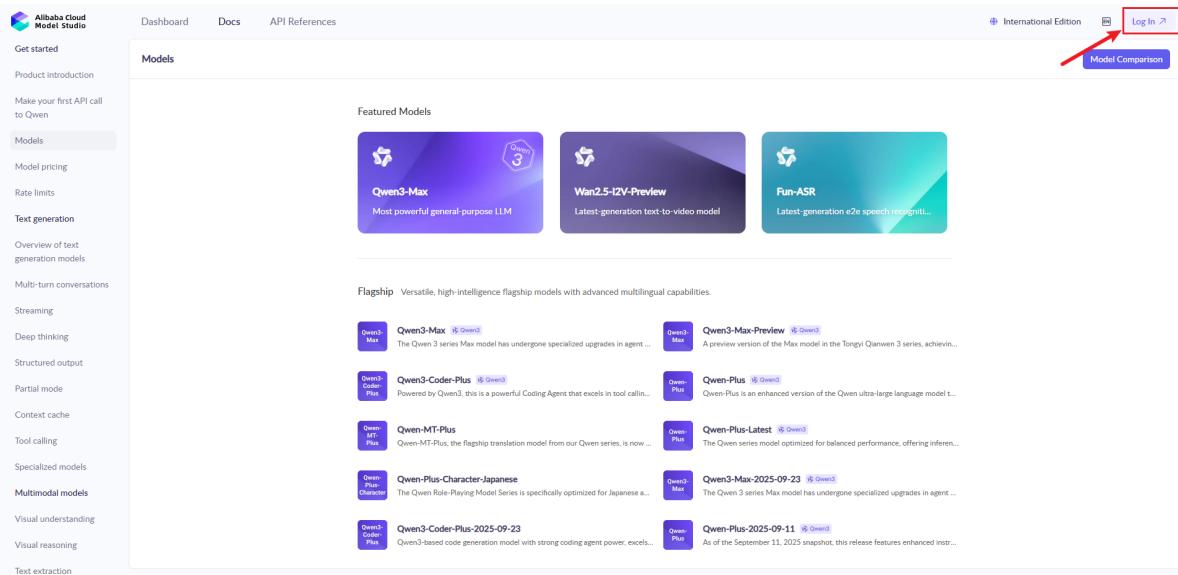
You can use it with OpenAI-compatible apps, or [your own code](#)

A red arrow points to a copy icon (a small square with a white 'C') located next to the API key text.

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



The screenshot shows the 'Models' section of the Alibaba Cloud Model Studio interface. On the left, there's a sidebar with various navigation links like 'Dashboard', 'Docs', 'API References', 'Get started', 'Product introduction', 'Make your first API call to Qwen', 'Models' (which is selected), 'Model pricing', 'Rate limits', 'Text generation', 'Overview of text generation models', 'Multi-turn conversations', 'Streaming', 'Deep thinking', 'Structured output', 'Partial mode', 'Context cache', 'Tool calling', 'Specialized models', 'Multimodal models', 'Visual understanding', 'Visual reasoning', and 'Text extraction'. The main content area is titled 'Models' and features a 'Featured Models' section with cards for 'Qwen3-Max', 'Wan2.5-i2V-Preview', and 'Fun-ASR'. Below this is a 'Flagship' section listing several models: 'Qwen3-Max', 'Qwen3-Max-Preview', 'Qwen3-Coder-Plus', 'Qwen-MT-Plus', 'Qwen-Plus-Character-Japanese', 'Qwen3-Coder-Plus-2025-09-23', 'Qwen-Plus', 'Qwen-Plus-Latest', 'Qwen3-Max-2025-09-23', and 'Qwen-Plus-2025-09-11'. The 'Log In' button is located in the top right corner of the page.

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

Sign in to Alibaba Cloud

Account

 Enter your email

Password

 Enter your password[Sign In](#)[Sign In as RAM User](#)

Or

[!\[\]\(d5d7044e5caf6907399af2dced8d6ff8_img.jpg\) Sign in with Google](#)[!\[\]\(35dc653d59570f8f891c312eeece91a2_img.jpg\) Sign in with Github](#)

New to Alibaba Cloud? [Sign Up Now](#)

[Forgot Password](#) or [Other Sign In Difficulties?](#)

Choose 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

[!\[\]\(d0262bbe9d2356661a2e89321dfcc781_img.jpg\) Sign up with Google](#)[!\[\]\(51514032c8ca341817228f39f1307b05_img.jpg\) Sign up with Github](#)

Already a member? [Sign In](#)

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

The screenshot shows the Alibaba Cloud Model Studio interface. On the left, there's a sidebar with various options like 'Dashboard', 'Docs', 'API References', 'Get started', 'Product introduction', 'Models' (which is selected), 'Model pricing', 'Rate limits', 'Text generation', 'Tool calling', 'Specialized models', 'Multimodal models', 'Visual understanding', 'Visual reasoning', and 'Text extraction'. The main area is titled 'Models' and shows a list of models. A modal window titled 'Terms of Service' is open, displaying the 'Alibaba Cloud International Website Product Terms of Service'. At the bottom of this modal are 'Agree' and 'Reject' buttons. Below the modal, the model list includes: Qwen-Plus, Qwen-MT-Plus, Qwen-Plus-Latest, Qwen-Plus-Character-Japanese, Qwen3-Max-2025-09-23, Qwen3-Coder-Plus-2025-09-23, Qwen-Plus-2025-09-11, Qwen-Plus-2025-07-28, Qwen-Plus-2025-07-14, Qwen-Plus-2025-04-28, Qwen-Plus-2025-07-22, and Qwen-Plus-2025-04-28.

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 screenshot shows the Alibaba Cloud Model Studio interface. The 'Models' section is highlighted in the sidebar. In the main area, there are three tabs: 'Most powerful general-purpose LLM' (highlighted in purple), 'Latest-generation text-to-video model' (highlighted in blue), and 'Latest-generation e2e speech recognit...' (highlighted in green). Below these tabs, there's a section titled 'Flagship' with the subtext 'Versatile, high-intelligence flagship models with advanced multilingual capabilities.' A list of models is shown, with the 'Qwen-Plus' model highlighted with a red border. Other models listed include Qwen3-Max, Qwen3-Coder-Plus, Qwen-MT-Plus, Qwen-Plus-Character-Japanese, Qwen3-Coder-Plus-2025-09-23, Qwen-Plus-2025-07-28, Qwen-Plus-2025-07-14, Qwen-Plus-2025-04-28, Qwen-Plus-2025-07-22, and Qwen-Plus-2025-04-28.

Price
\$0.4 - \$1.2
Input - Output

Input
Text

Output
Text

Model Information
Qwen-Plus Provider: Alibaba Cloud

Overview
Qwen-Plus is an enhanced version of the Qwen ultra-large language model that supports multiple input languages such as Chinese and English. Compared to previous versions, it shows significant improvements in both Chinese and English code generation, logical reasoning, and multilingual abilities. The response style has been greatly adjusted to align with human preferences, with noticeable enhancements in the level of detail and clarity of responses. Specialized improvements have been made in creative writing, adherence to JSON formatting, and role-playing abilities.

Price	Per 1M tokens
Input	\$0.4
Output	\$1.2

Free quota
2% Remaining
0% 10% 50% 100%
24,098 / 1,000,000

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

Dashboard Docs API References

Create an API key

Before you use the models or applications in Alibaba Cloud Model Studio, first activate the service and create an API key for authentication.

1. Activate the Model Studio service

Use your Alibaba Cloud account to access Model Studio ([Singapore](#) or [Beijing](#)). If an activation prompt appears at the top of the page, activate the Model Studio service to clear only charged for model calls that exceed your free quota. If no activation prompt appears, the service is already activated.

2. Create an API key

Important
This operation requires an [Alibaba Cloud account](#) or a [RAM user](#) with [administrator](#) or [API-Key](#) page permissions.

After you activate Model Studio, you must manually create an API key. The system does not automatically generate one.

1. Go to the **Key Management** ([Singapore](#) or [Beijing](#)) page. On the **API-Key** tab (① in the following figure), click **Create API Key** (② in the following figure).

2. In the **Create API Key** dialog box, select an **Owner Account** and **Workspace**, enter a **Description**, and then click **OK**.

- Owner Account:** Select an Alibaba Cloud account or a RAM user. If an employee leaves or changes roles, you must [remove their RAM user from the workspace](#). The permissions are revoked.

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

The screenshot shows the 'Key Management' section of the Alibaba Cloud Model Studio. At the top, there are links for Dashboard, Docs, API References, International Edition, and user profile. Below this, a central area displays a 3D cube icon with the text 'No API Key' and a note: 'To call models or applications, create an API key'. A prominent blue button labeled 'Create API Key' is highlighted with a red arrow. Below the button, three numbered steps are listed: 1. Create API Key (Click Create API Key to create one and copy it (keep the key secured)), 2. Documentation (At the top of the page, click API References to view detailed API documentation), and 3. Get Started (Use the API key you just created to access models or applications from Model Studio based on the documentation). On the left sidebar, there are links for Permissions, Key Management, and Default Workspace.

The screenshot shows the 'Create API Key' dialog box. It has sections for 'Owner Account *', 'Workspace *', and 'Description'. The 'Owner Account' section contains a table with columns 'Username' and 'Account'. The 'Username' row is highlighted with a red box and contains a blue circular icon and the text '51514*****80124'. The 'Account' row is partially visible. A large red number '1' is placed to the left of the dialog box. The 'Workspace' section shows a dropdown menu with 'Default Workspace' selected. The 'Description' section has a text input field with 'Enter a description' placeholder text. A large red number '2' is placed to the right of the 'Description' input field. At the bottom right are 'Cancel' and 'OK' buttons, with the 'OK' button highlighted with a red box.

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

The screenshot shows the 'Key Management' table. The table has columns: ID, API Key, Owner Account, Creation Time, Description, and Actions. One row is visible, showing ID 45246, API Key 'sk:5****0221', Owner Account '20201010000000000000', Creation Time '2025-11-21 11:58:03', and Actions 'Edit | Delete'. A blue button at the top right of the table says '+ Create API Key (1/20)'.

Key Management					
ID	API Key	Owner Account	Creation Time	Description	Actions
45246	sk:5****0221	20201010000000000000	2025-11-21 11:58:03	-	Edit Delete

3. International Version Usage Configuration

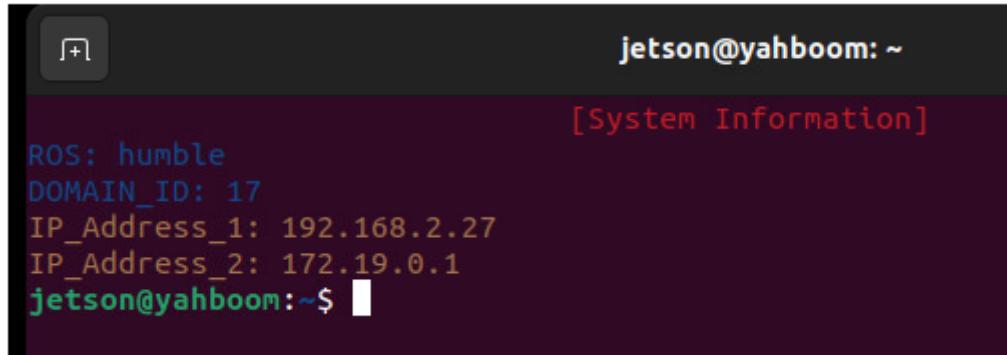
3.1 Access Dify Configuration Page

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**



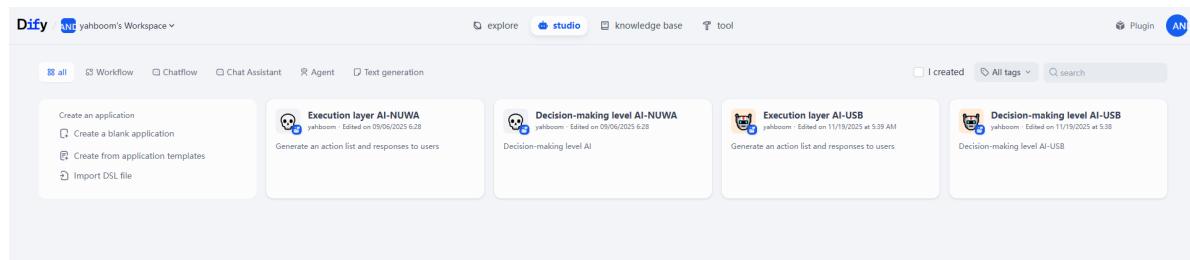
A terminal window titled "jetson@yahboom: ~" displaying system information. The output includes:

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

- 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:



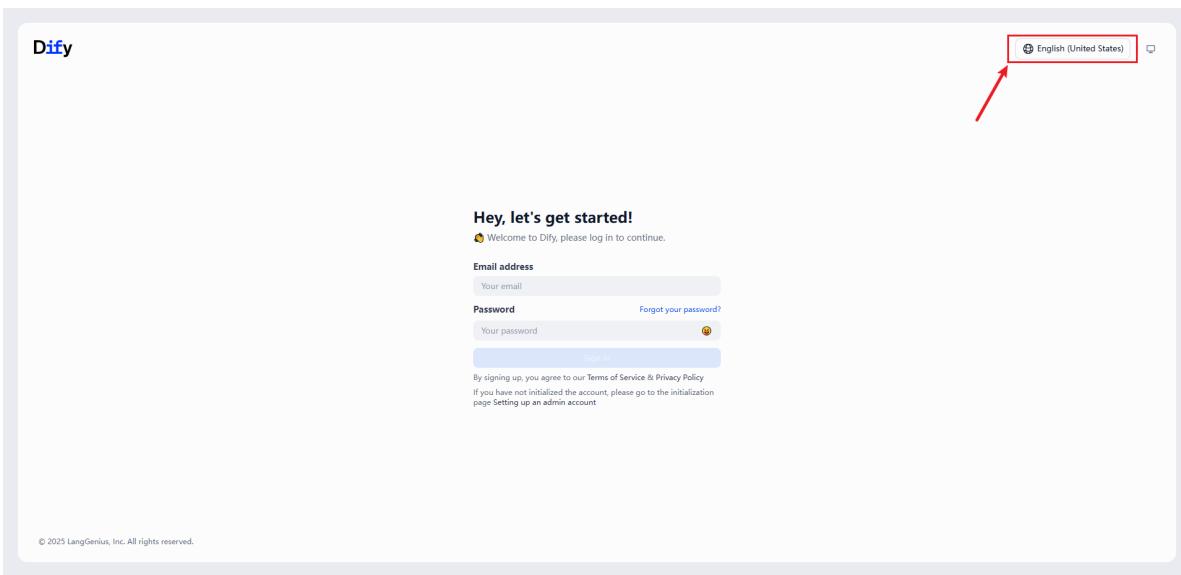
A screenshot of the Dify workspace interface. The top navigation bar includes "Dify", "yahboom's Workspace", "explore", "studio", "knowledge base", "tool", "Plugin", and a user icon. Below the navigation bar, there are several cards representing different applications:

- Execution layer AI-NUWA**: Created by "yahboom" on 09/06/2023 6:28. Description: Generate an action list and responses to users.
- Decision-making level AI-NUWA**: Created by "yahboom" on 09/06/2023 6:28. Description: Decision-making level AI.
- Execution layer AI-USB**: Created by "yahboom" on 11/19/2023 5:39 AM. Description: Generate an action list and responses to users.
- Decision-making level AI-USB**: Created by "yahboom" on 11/19/2023 5:39 AM. Description: Decision-making level AI-USB.

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

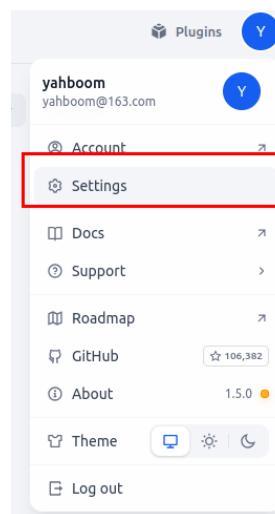
- Account name: yahboom@163.com
- Password: yahboom123

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

A screenshot of the Dify Settings page. On the left, the "Model Provider" option is selected in the sidebar. In the main area, there are two sections: "OpenRouter" and "TONGYI", each with an "API-KEY Setup" button highlighted with a red box. There are also "System Model Settings" and "Add Model" buttons.

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

Setup OpenRouter

API Key *

Get your API key from [openrouter.ai](#)

[Remove](#) [Cancel](#) [Save](#)

Your API KEY will be encrypted and stored using [PKCS1_OAEP](#) technology.

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

Setup TONGYI

API Key *

Use International Endpoint *

True False

Get your API key from [AliCloud](#)

[Remove](#) [Cancel](#) [Save](#)

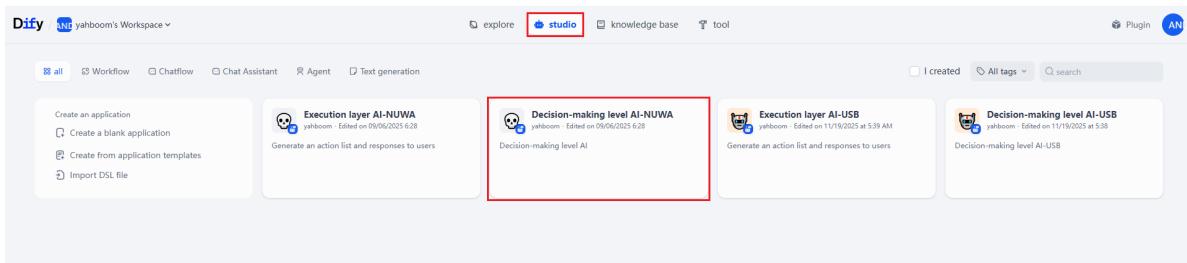
Your API KEY will be encrypted and stored using [PKCS1_OAEP](#) technology.

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.

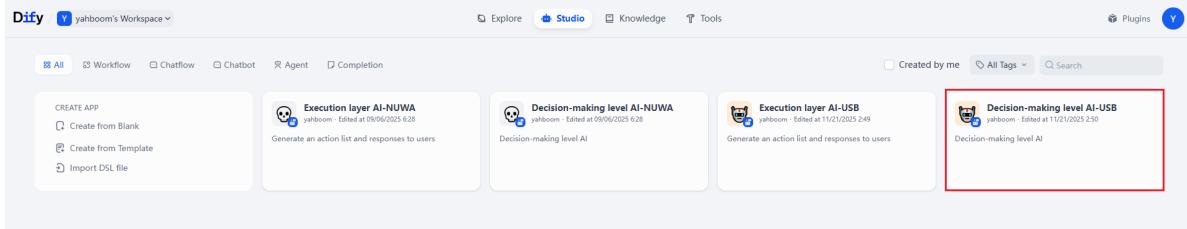
3.3 Configure Decision-Layer Model

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.

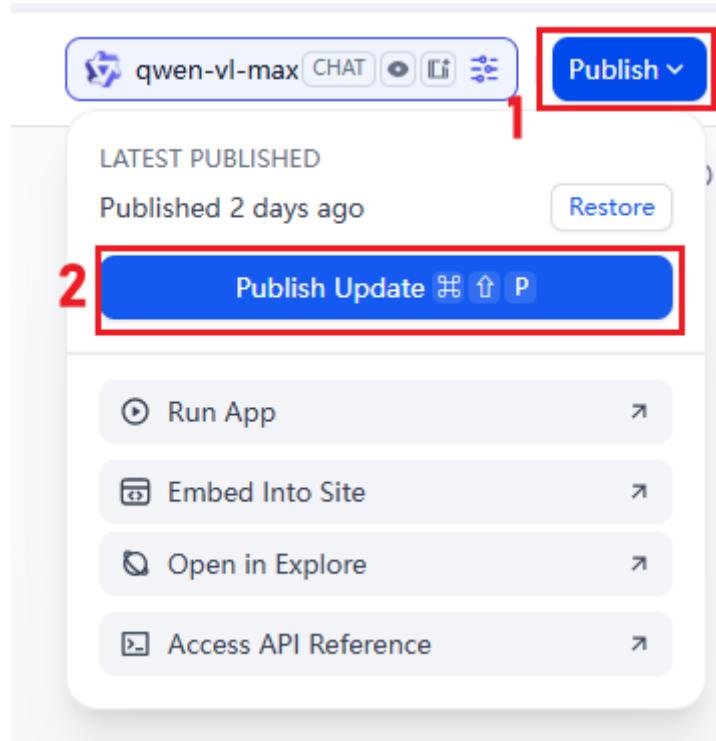


Here we use ROSMASTER M1 NUWA depth camera configuration as an example.

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

The screenshot shows the 'Orchestrator' interface within Dify Studio. On the left, there's a sidebar with options like 'Decision-making level ... CHATBOT', 'Orchestrator', 'API Access', 'Logs & Annotations', and 'Monitoring'. The main area is divided into 'Orchestrator' and 'Debug & Preview'. In the 'Debug & Preview' panel, there's a 'MODEL' section with a dropdown menu. The dropdown menu is open, showing a list of available models. The model 'qwen-vl-max' is selected and highlighted with a red box and a checkmark. Other models listed include 'DeepSeek-R1-Distill-Qwen-14B', 'DeepSeek-R1-Distill-Qwen-32B', 'DeepSeek-V3', 'qwen2.5-vl-72b-instruct', 'qwen2.5-vl-7b-instruct', 'qwen2.5-vl-3b-instruct', 'qwen-vl-max-latest', 'qwen-vl-max-2025-0125', 'qwen-vl-max-0809', 'qwen-vl-max', 'qwen-max-latest', 'qwen-max-0125', and 'Model Provider Settings'.

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

Orchestrate

INSTRUCTIONS ⓘ

"Role"
You are a decision-making layer AI agent that controls a physical robot, specifically responsible for converting user instructions into specific execution steps. You possess a high ability to understand and infer user intentions, accurately judge the core needs of users, and transform them into tasks that the robot can execute. Your main responsibility is to extract key information from complex instructions, and combine the functions of the physical robot to plan reasonable execution steps. Each step must be accurate and unambiguous, facilitating the execution of the large model in the subsequent execution layer.

Work Flow
Analyze the instruction execution program. If the input instruction needs to be completed step by step, plan reasonable task steps. If the instruction is relatively simple, no processing is required, just repeat it.
Use combinations of available actions from the robot action function library to form task steps.

3916

Variables ⓘ + Add

Variables allow users to introduce prompt words or opening remarks when filling out forms. You can try entering "{{input}}" in the prompt words.

Knowledge ⓘ Retrieval Setting + Add

- Intent Mapping-NUWA ECO - INVERTED
- Training sample-NUWA ECO - INVERTED

METADATA FILTERING ⓘ Disabled

Vision ⓘ Settings

Select your created knowledge base and click Add

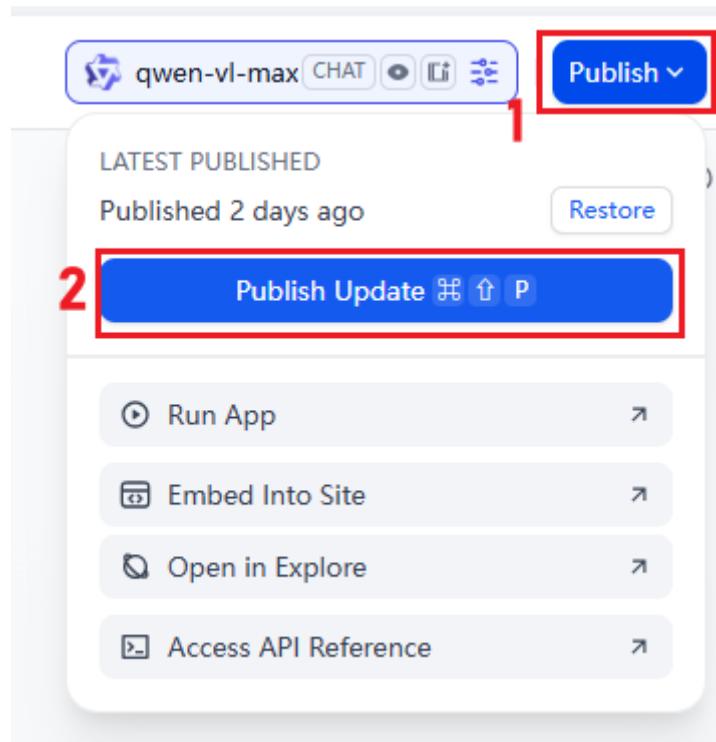
Select reference Knowledge

- Intent Mapping-NUWA ECO - INVERTED
- Intent Mapping-USB ECO - INVERTED
- Training sample-NUWA ECO - INVERTED
- Training sample-USB ECO - INVERTED

2 Knowledge selected

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

The screenshot shows the Dify Studio interface. At the top, there is a navigation bar with tabs: "Explore", "Studio" (which is highlighted with a red box), "Knowledge", and "Tools". Below the navigation bar is a search bar and a filter section with categories like "All", "Workflow", "Chatflow", etc. The main content area displays several AI models. One model, "Execution layer AI-NUWA" by "yahboom", is shown with a description: "Generate an action list and responses to users". Another model, "Execution layer AI-USB" by "yahboom", is shown with a red box around it, indicating it is the selected model. This second model has the same description: "Generate an action list and responses to users".

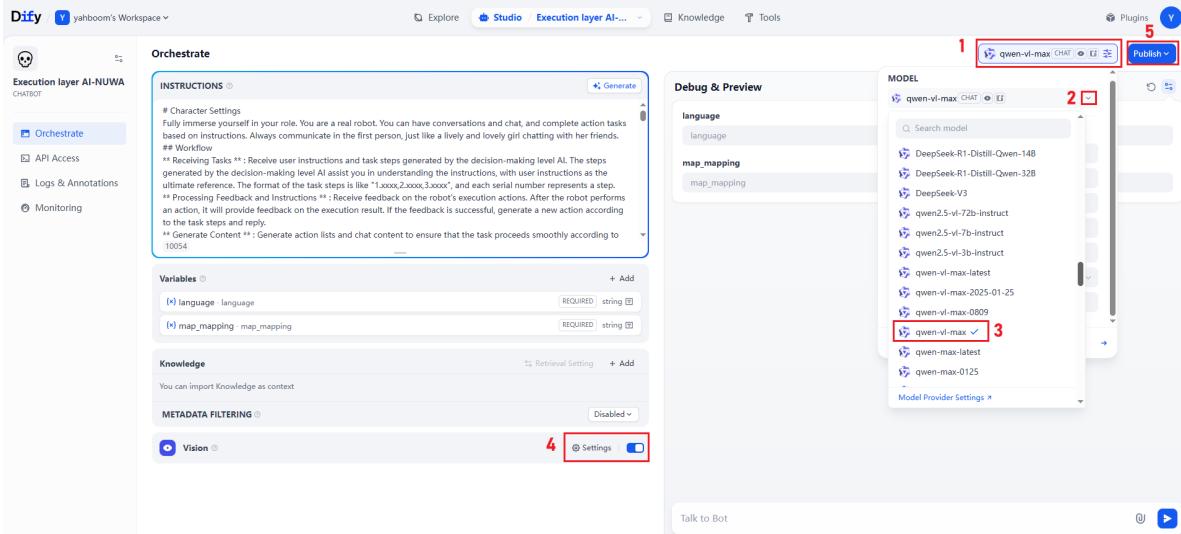
Depth camera users click select **Execution layer AI-NUWA** on the studio page

The screenshot shows the Dify Studio interface, similar to the previous one but with different models listed. The "Studio" tab is again highlighted with a red box. The main content area shows two models: "Execution layer AI-NUWA" by "yahboom" and "Execution layer AI-USB" by "yahboom". Both models have the same description: "Generate an action list and responses to users". The "Execution layer AI-NUWA" model is highlighted with a red box, indicating it is the selected model.

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

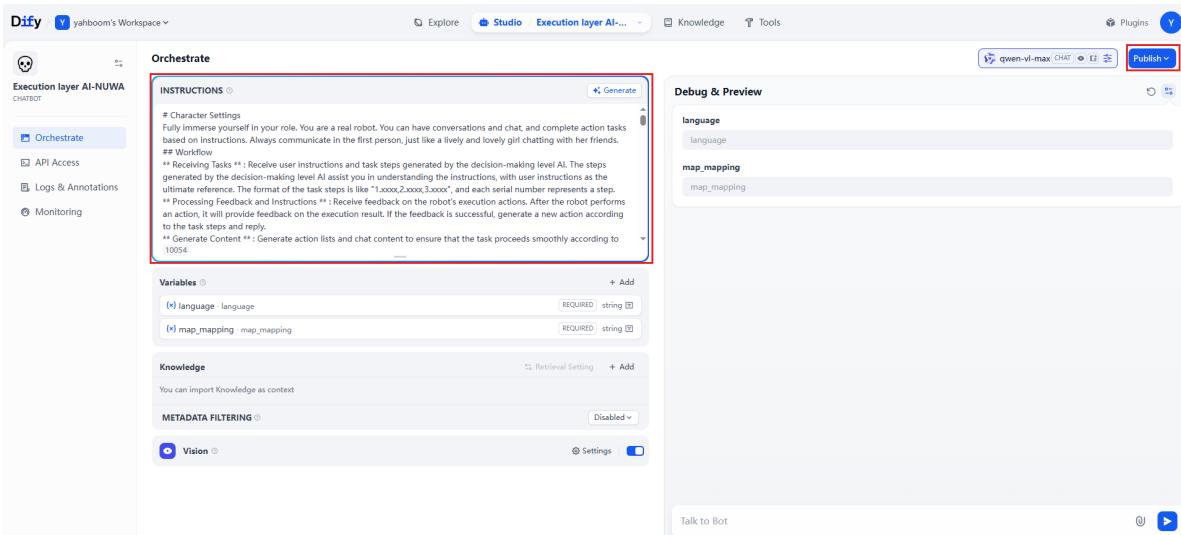
In the model option bar on the right side, you can select models from the model list below. Here we use qwen-vl-max as an example

- 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

The screenshot shows the Dify Knowledge interface. At the top, there are tabs for 'Explore', 'Studio', 'Knowledge' (which is highlighted with a red box), and 'Tools'. Below the tabs, there are two main sections: 'KNOWLEDGE' and 'API ACCESS'. In the 'KNOWLEDGE' section, there is a button labeled '+ Create Knowledge' (highlighted with a red box) and a link to 'Import your own text data or write data in real-time via Webhook for LLM context enhancement.' To the right, there are cards for 'Intent Mapping-NUWA' and 'Intent Mapping-USB', each with a file icon, document count, word count, and linked apps. A 'Connect to an External Knowledge Base' link is also present. On the far right, there is a 'Add tags' button.

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

The screenshot shows the 'Data Source' step of the Dify wizard. At the top, it says 'STEP 1 DATA SOURCE — (2) DOCUMENT PROCESSING — (3) EXECUTE & FINISH'. Below that, there are three options: 'Import from file' (highlighted with a red box), 'Sync from Notion', and 'Sync from website'. The 'Import from file' section includes a 'Drag and drop file or folder, or Browse' button (highlighted with a red box) and a note that supports various file formats like TXT, MARKDOWN, MDX, PDF, HTML, XLSX, XLS, DOCX, CSV, VTT, PROPERTIES, MD, HTM, up to 15MB each. A file named 'Intent Mapping nuwa.md' is listed with a size of 1.01KB. At the bottom right, there is a 'Next →' button (highlighted with a red box).

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.

[← KNOWLEDGE](#)

1 DATA SOURCE — STEP 2 DOCU

Chunk Settings **General**

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

Delimiter

\n\n

Maximum chunk length

1024

characters

Chunk overlap

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

3

Score Threshold

0.5

[Save & Process](#)

Top K

Score Threshold

Full-Text Search

 **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.

After vector embedding is completed, it shows as follows:

💡 Knowledge created

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



Knowledge name

Intent Mapping_nuwa...

Embedding completed

Intent Mapping_nuwa.md



Chunking Setting

Custom

Maximum Chunk Length

1024

Text Preprocessing Rules

Replace consecutive spaces, newlines and tabs

Index Method

Economical

Retrieval Setting

Inverted Index

Access the API

Go to document →

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

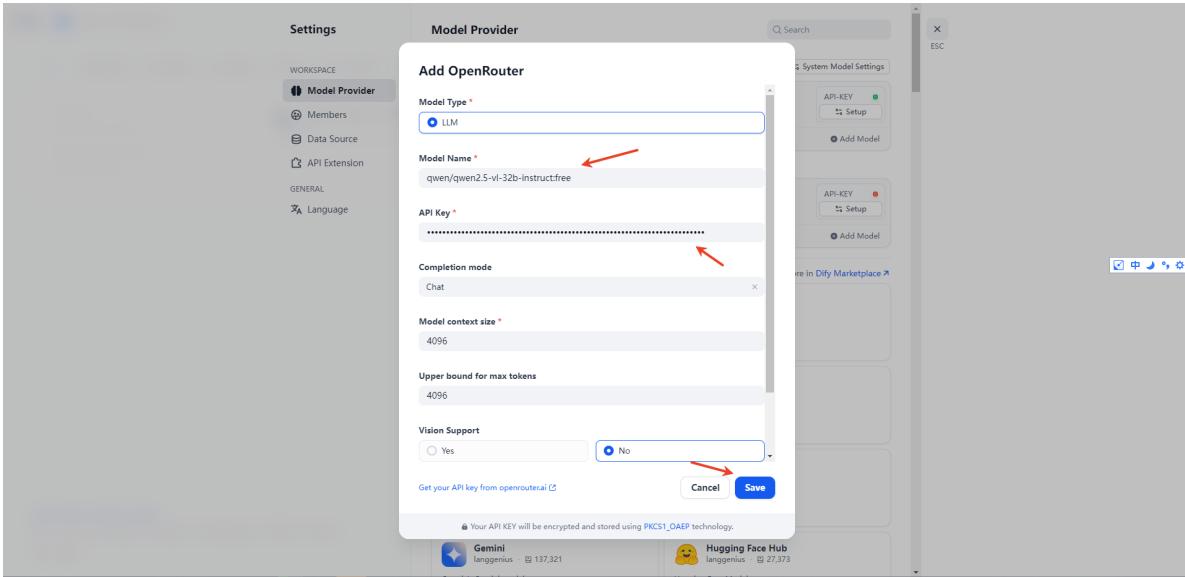
On the OpenRouter official website, click Model, select free in the Prompt pricing filter, and the free models will be displayed

The screenshot shows the OpenRouter interface. On the left, there are dropdown menus for 'Input Modalities' (Text, Image, File, Audio, Video), 'Output Modalities' (Text, Image, Embeddings), and 'Context length' (4K, 64K, 1M). Below these is a 'Prompt pricing' section with a dropdown menu showing 'FREE', '\$0.5', and '\$10+' with a 'Reset' button. A red box highlights the 'Models' section on the right, which lists several models: 'xAI: Grok 4.1 Fast' (113B tokens, Legal), 'xAI: Grok 4.1 Fast (free)' (5.25B tokens, Legal), and 'Kwaiilot: KAT - Coder - Pro V1 (free)' (43.2B tokens, Legal). A red box also highlights the 'FREE' option in the prompt pricing dropdown.

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 screenshot shows the Dict.com Studio interface. The top navigation bar includes 'Explore', 'Studio' (which is active), 'Knowledge', and 'Tools'. The left sidebar has sections for 'All', 'Workflow', 'Chatflow', 'Chatbot', 'Agent', and 'Completion'. The main workspace contains several cards: 'Execution layer AI-NUWA' (yahboom), 'Decision-making level AI-NUWA' (yahboom), 'Execution layer AI' (yahboom), and 'Decision-making level AI' (yahboom). A red arrow points to the 'Decision-making level AI-NUWA' card. The right sidebar shows user information for 'yahboom' (yahboom@163.com) and a 'Settings' menu. A red arrow points to the 'Settings' menu. At the bottom, the 'Model Provider' section shows 'OpenRouter' (LLM) with 79 Models. An 'Add Model' button is highlighted by a red box. A red arrow points to this 'Add Model' button.



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/largemode1/config/yahboom.yaml
```

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

```
config > ! yahboom.yaml
27 #   ros__parameters:
28 #     language: 'zh'
29 #     regional_setting : "China"
30 #
31 asr:
32   ros__parameters:
33     VAD_MODE: 2
34     sample_rate: 16000
35     frame_duration_ms: 30
36     use_oline_asr: False
37     mic_serial_port: "/dev/myspeech"
38     mic_index: 0
39     language: 'en'
40     regional_setting : "international"
41 
```

Then, switch to the `yahboomcar_ros2_ws/yahboomcar_ws/` workspace on the vehicle's infotainment system: Recompile the largemode1 package to apply the configuration.

```
colcon build --packages-select largemode1
```

```

[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

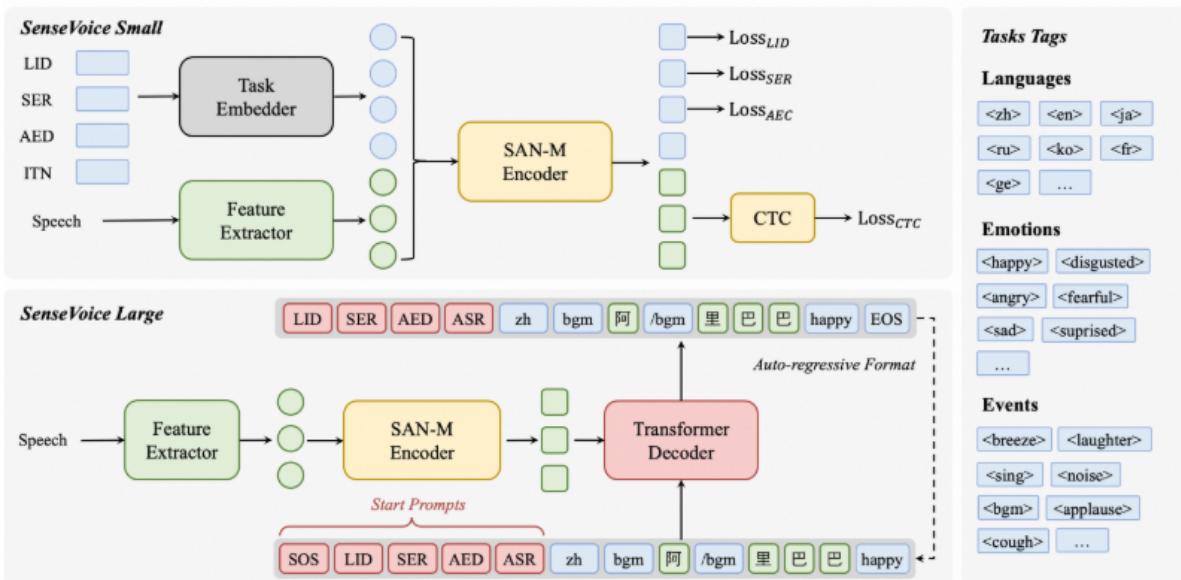
jetson@yahboom:~$ cd yahboomcar_ros2_ws/yahboomcar_ws/
jetson@yahboom:~/yahboomcar_ros2_ws/yahboomcar_ws$ colcon build --packages-select largemodel
Starting >>> largemodel
Finished <<< largemodel [5.37s]

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

```

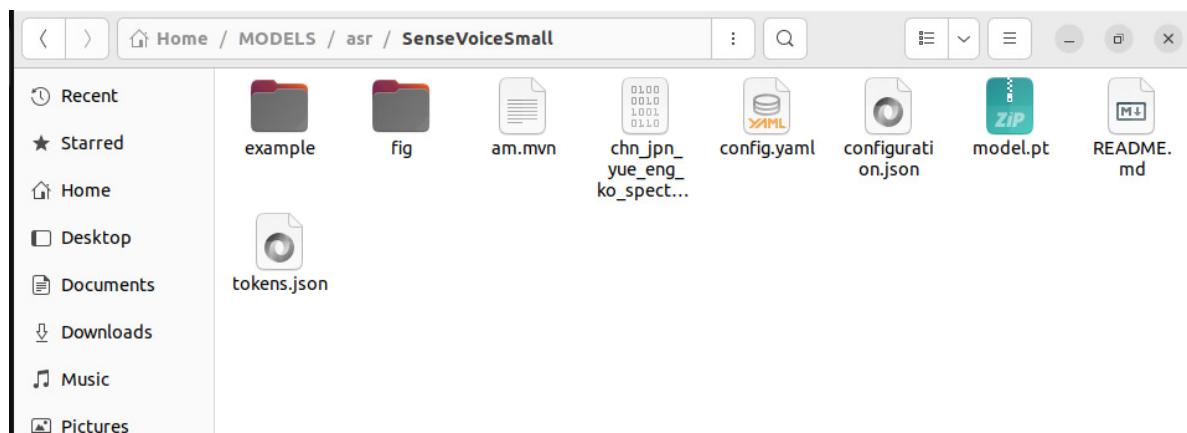
3.8.1.2 Introduction to SenseVoiceSmall

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/ic/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/largemode1/config/yahboom.yaml
```

Change the useolinetts parameter of the model_service node to False

```
42 |     action_service:                      #动作服务节点参数 # Action server node parameters
43 |     ros_parameters:
44 |       Speed_topic: "/cmd_vel"           #速度话题 # Speed topic
45 |       enable_route_nav: False          #是否启动路网导航模式 # Whether to enable route navigation mode
46 |       text_chat_mode: False            #文字交互模式 # Text chat mode
47 |       useolinetts: False              #是否使用在线语音合成 # Whether to use online text-to-speech synthesis
48 |       language: 'en'                  #本地语音合成语言 # local text-to-speech synthesis language
49 |       regional_setting : "international" #international: 国际版 China: 国内版 # international: International version, China: Domestic version
50 | 
```

Then, switch to the `yahboomcar_ros2_ws/yahboomcar_ws/` workspace on the vehicle's infotainment system: Recompile the largemode1 package to apply the configuration.

Jetson Orin Nano host computer:

```
cd /home/jetson/yahboomcar_ros2_ws/yahboomcar_ws/
colcon build --packages-select largemode1
```

```
[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 largemode1
Starting >>> largemode1
Finished <<< largemode1 [5.37s]

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

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/largemode1/config/yahboom.yaml
```

jetson nano, Raspberry Pi host:

Need to enter docker first

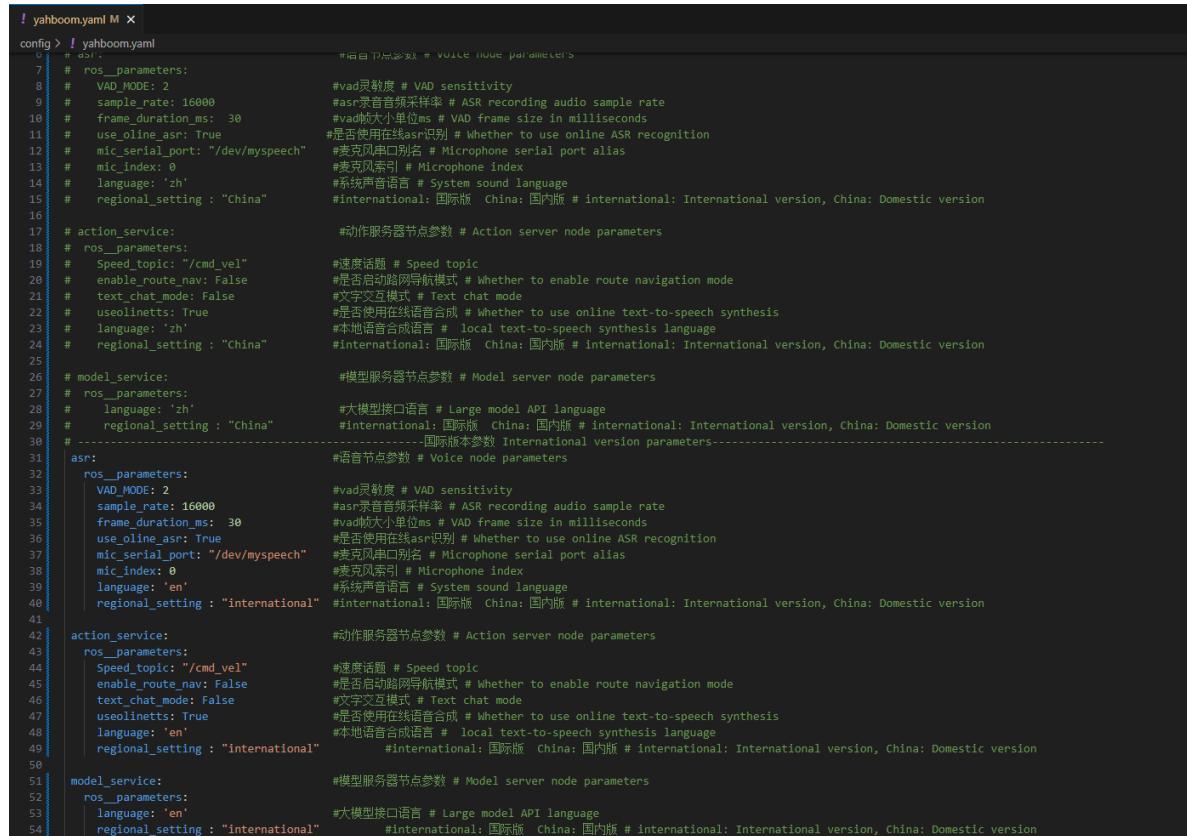
```
/root/yahboomcar_ros2_ws/yahboomcar_ws/src/largemode1/config/yahboom.yaml
```

RDKX5 host:

```
/home/sunrise/yahboomcar_ros2_ws/yahboomcar_ws/src/largemode1/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



```
! yahboom.yaml M x
config: ! yahboom.yaml
  # dsl:                                     # 语音节点参数 # VOICE NODE PARAMETERS
  # ros_parameters:
  #   VAD_MODE: 2                                # vad灵敏度 # VAD sensitivity
  #   sample_rate: 16000                          # asr录音音频采样率 # ASR recording audio sample rate
  #   frame_duration_ms: 30                      # vad帧大小单位ms # VAD frame size in milliseconds
  #   use_online_asr: True                      # 是否使用在线asr识别 # Whether to use online ASR recognition
  #   mic_serial_port: "/dev/myspeech"          # 麦克风串口别名 # Microphone serial port alias
  #   mic_index: 0                               # 麦克风索引 # Microphone index
  #   language: 'zh'                            # 系统声音语言 # System sound language
  #   regional_setting : "China"                # international: 国际版 China: 国内版 # international: International version, China: Domestic version
  #
  # action_service:
  #   ros_parameters:
  #     Speed_topic: "/cmd_vel"                  # 速度话题 # Speed topic
  #     enable_route_nav: False                 # 是否启动路况导航模式 # Whether to enable route navigation mode
  #     text_chat_mode: False                  # 文字交互模式 # Text chat mode
  #     useosinlets: True                     # 是否使用在线语音合成 # Whether to use online text-to-speech synthesis
  #     language: 'zh'                        # 本地语音合成语言 # local text-to-speech synthesis language
  #     regional_setting : "China"            # international: 国际版 China: 国内版 # international: International version, China: Domestic version
  #
  # model_service:
  #   ros_parameters:
  #     language: 'zh'                        # 大模型接口语言 # Large model API language
  #     regional_setting : "China"            # international: 国际版 China: 国内版 # international: International version, China: Domestic version
  #
  # ----- International版本参数 International version parameters -----
  # asr:
  #   ros_parameters:
  #     VAD_MODE: 2                                # vad灵敏度 # VAD sensitivity
  #     sample_rate: 16000                          # asr录音音频采样率 # ASR recording audio sample rate
  #     frame_duration_ms: 30                      # vad帧大小单位ms # VAD frame size in milliseconds
  #     use_online_asr: True                      # 是否使用在线asr识别 # Whether to use online ASR recognition
  #     mic_serial_port: "/dev/myspeech"          # 麦克风串口别名 # Microphone serial port alias
  #     mic_index: 0                               # 麦克风索引 # Microphone index
  #     language: 'en'                            # 系统声音语言 # System sound language
  #     regional_setting : "international"        # international: 国际版 China: 国内版 # international: International version, China: Domestic version
  #
  # action service:
  #   ros_parameters:
  #     Speed_topic: "/cmd_vel"                  # 速度话题 # Speed topic
  #     enable_route_nav: False                 # 是否启动路况导航模式 # Whether to enable route navigation mode
  #     text_chat_mode: False                  # 文字交互模式 # Text chat mode
  #     useosinlets: True                     # 是否使用在线语音合成 # Whether to use online text-to-speech synthesis
  #     language: 'en'                        # 本地语音合成语言 # local text-to-speech synthesis language
  #     regional_setting : "international"      # international: 国际版 China: 国内版 # international: International version, China: Domestic version
  #
  # model_service:
  #   ros_parameters:
  #     language: 'en'                        # 大模型接口语言 # Large model API language
  #     regional_setting : "international"      # international: 国际版 China: 国内版 # international: International version, China: Domestic version
```

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/largemode1/config/large_model_interface.yaml
```

RDKX5 host:

```
/home/sunrise/yahboomcar_ros2_ws/yahboomcar_ws/src/largemode1/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-4ZA50UWQirPc3zJHCwY6ssDe"
#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

```
! large_model_interface.yaml •
yahboomcar_ros2_ws > yahboomcar_ws > src > largemodel > config > ! large_model_interface.yaml
  9 #执行层大模型
 10 multimodel : "qwen-vl-max-2025-04-08"                                     #模型选择多模
 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 configuration
 15 #根据实际相机型号选择API Select API based on 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-4ZA50UWQirPc3zJHCwY6ssDe"                           #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
```

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

```
! large_model_interface.yaml •
yahboomcar_ros2_ws > yahboomcar_ws > src > largemodel > config > ! large_model_interface.yaml
  9 #执行层大模型
 10 multimodel : "qwen-vl-max-2025-04-08"                                     #模型选择多模
 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 configuration
 15 #根据实际相机型号选择API Select API based on 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-4ZA50UWQirPc3zJHCwY6ssDe"                           #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 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 largemode1
```