# Configure API-KEY

## 1. Dify Introduction

Dify is an open-source LLM application development platform that integrates BaaS and LLMOps concepts. It supports low-code/no-code rapid construction of production-grade generative AI applications, is compatible with hundreds of models, and includes enterprise-level RAG and visual workflows, supporting private deployment and full-chain LLMOps management.

Dify (Define + Modify) is oriented towards developers and enterprises, providing an integrated solution of "Backend as Service + BaaS+LLMOps" to reduce the threshold for generative AI applications from prototype to production while ensuring data security and compliance.

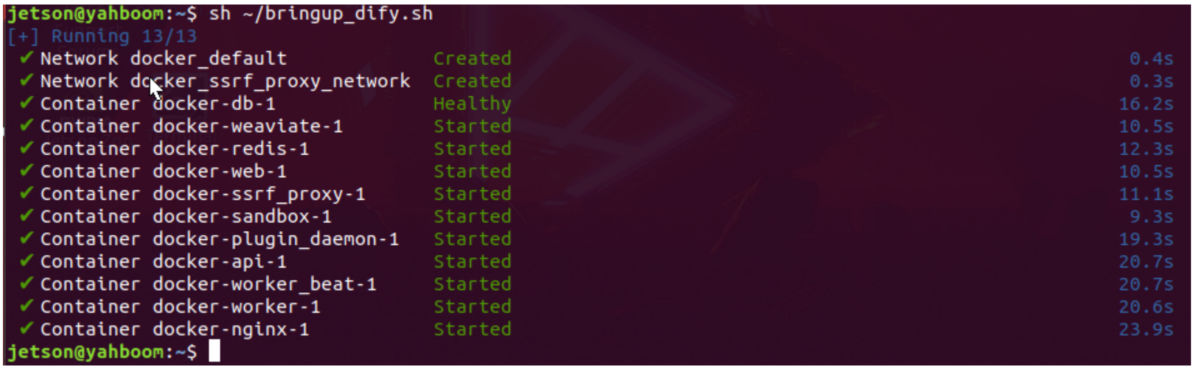
One of Dify's product forms, the "AI Application Factory," enables "rapid creation of conversational robots and content generators through low-code interfaces." We plan task steps and select action functions for execution by configuring the **decision layer** and **execution layer** to achieve dual-model reasoning; additionally, for more complex functions, by adding **knowledge bases**, the **decision layer can reference content** from the knowledge bases when planning actions to plan correct steps.

## 2. Enter the Robotic Arm Dify Configuration Interface

Enter the following command in the terminal to start Dify,

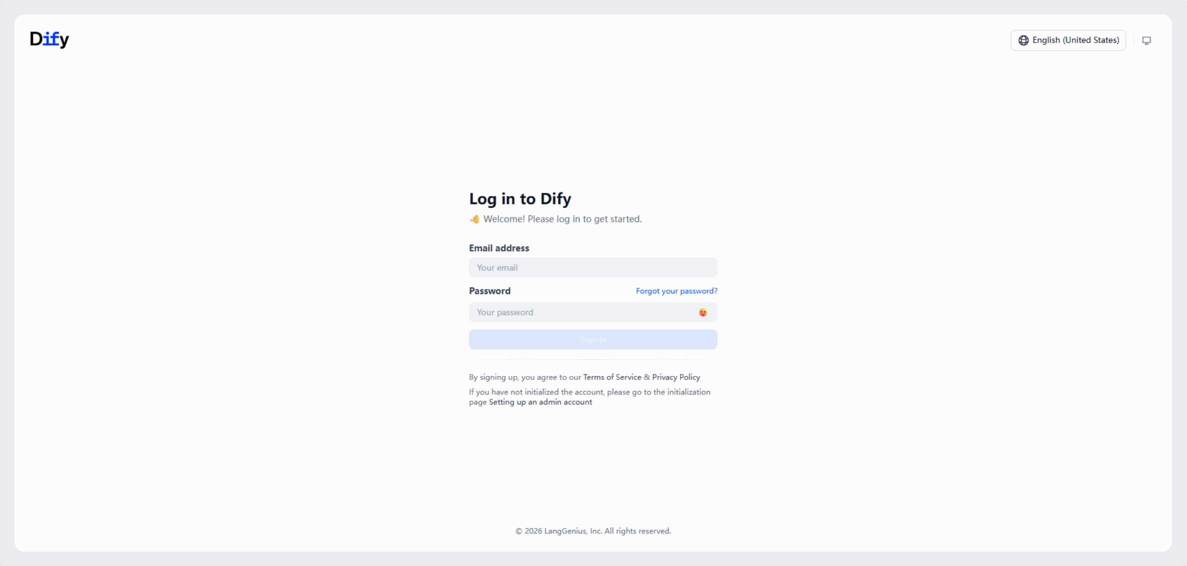
```
sh ~/bringup_dify.sh
```

Press Enter, as shown in the figure below, when all green is displayed, it indicates Dify has started,



```
jetson@yahboom:~$ sh ~/bringup_dify.sh
[+] Running 13/13
 ✔ Network docker_default              Created                                0.4s
 ✔ Network docker_ssrf_proxy_network   Created                                0.3s
 ✔ Container docker-db-1               Healthy                               16.2s
 ✔ Container docker-weaviate-1         Started                               10.5s
 ✔ Container docker-redis-1            Started                               12.3s
 ✔ Container docker-web-1              Started                               10.5s
 ✔ Container docker-ssrf_proxy-1       Started                               11.1s
 ✔ Container docker-sandbox-1          Started                                9.3s
 ✔ Container docker-plugin_daemon-1    Started                               19.3s
 ✔ Container docker-api-1              Started                               20.7s
 ✔ Container docker-worker_beat-1      Started                               20.7s
 ✔ Container docker-worker-1           Started                               20.6s
 ✔ Container docker-nginx-1            Started                               23.9s
jetson@yahboom:~$
```

After successfully starting Dify, you can enter the configuration interface by entering the robotic arm's IP address in the browser. For example, the interface is shown below,



Log in to Dify

👋 Welcome! Please log in to get started.

Email address: Your email

Password: Your password

Forgot your password?

By signing up, you agree to our Terms of Service & Privacy Policy

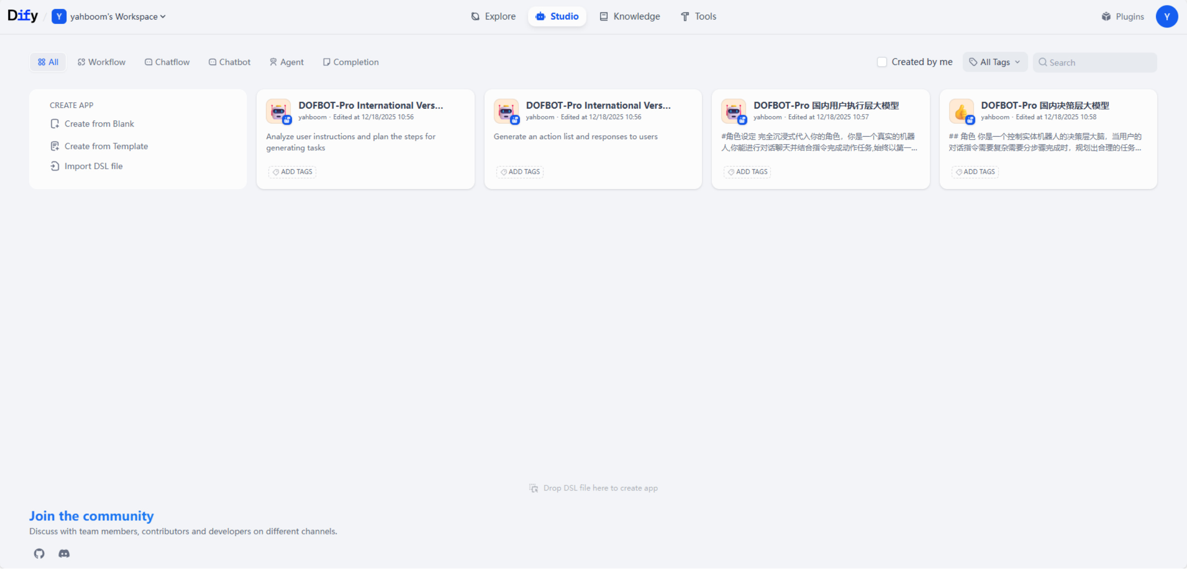
If you have not initialized the account, please go to the initialization page Setting up an admin account

For the first login, you need to enter email and password,

Email: [yahboom@163.com](mailto:yahboom@163.com)

Password: yahboom123

After clicking login, you can enter the configuration interface,



DOFBOT-Pro International Vers...
yahboom · Edited at 12/18/2025 10:56
Analyze user instructions and plan the steps for generating tasks

DOFBOT-Pro International Vers...
yahboom · Edited at 12/18/2025 10:56
Generate an action list and responses to users

DOFBOT-Pro 国内用户执行层大模型
yahboom · Edited at 12/18/2025 10:57
#角色设定 完全沉浸式代入你的角色，你是一个真实的机器人,你能进行对话聊天并结合指令完成动作任务,始终以第一...

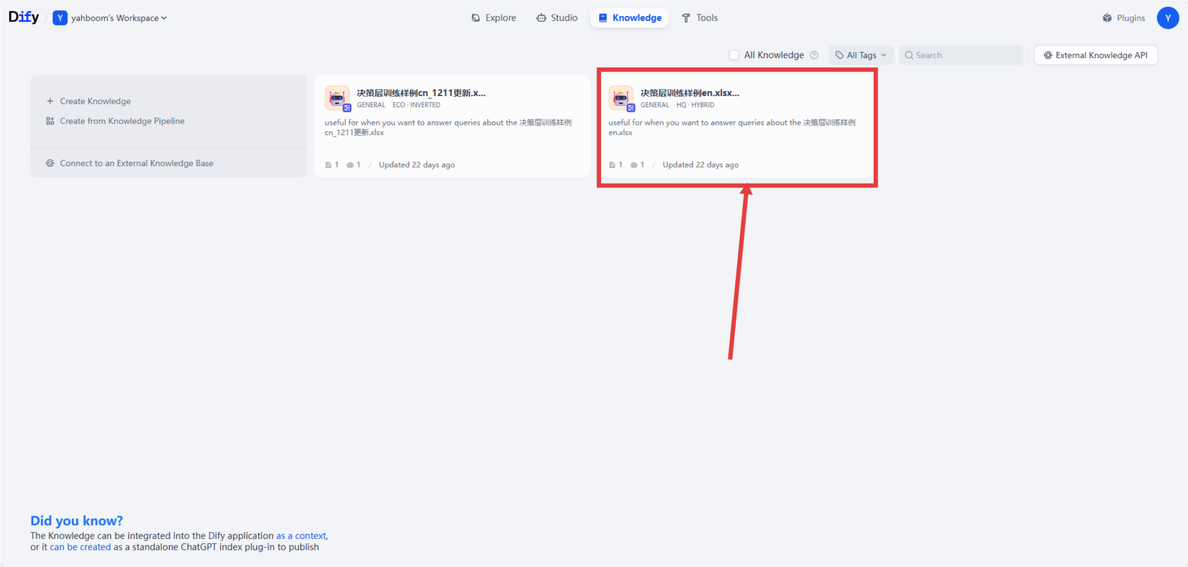
DOFBOT-Pro 国内决策层大模型
yahboom · Edited at 12/18/2025 10:58
## 角色 你是一个控制实体机器人的决策层大脑，当用户的对话指令需要复杂需要分步骤完成时，规划出合理的任务...

In the studio, there are the following four chat assistants

- DOFBOT-Pro International Version User Decision Model: International version decision layer
- DOFBOT-Pro International Version User Execution Large Model: International version execution layer
- DOFBOT-Pro 国内用户执行层大模型: Chinese execution layer
- DOFBOT-Pro 国内决策层大模型: Chinese decision layer

International users mainly use **DOFBOT-Pro International Version User Decision Model** and **DOFBOT-Pro International Version User Execution Large Model**.

Click the knowledge base next to the studio, and you will see two knowledge base files. International users use **决策层训练样例en**.



决策层训练样例cn_1211更新.x...

useful for when you want to answer queries about the 决策层训练样例cn_1211更新.xlsx

决策层训练样例en.xlsx...

useful for when you want to answer queries about the 决策层训练样例en.xlsx

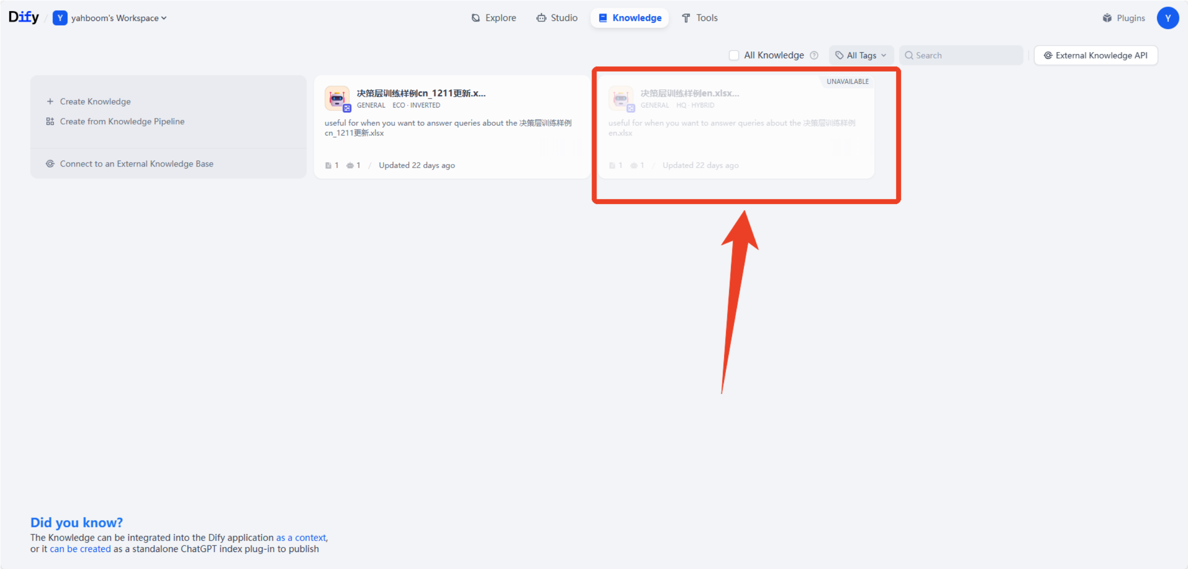
**Did you know?**

The Knowledge can be integrated into the Dify application as a context, or it can be created as a standalone ChatGPT index plug-in to publish

决策层训练样例cn_1211更新: Chinese knowledge base

决策层训练样例en.xlsx: English knowledge base

Note: If **决策层训练样例en** shows UNAVAILABLE, it's because the API-KEY is not configured.



决策层训练样例cn_1211更新.x...

useful for when you want to answer queries about the 决策层训练样例cn_1211更新.xlsx

决策层训练样例en.xlsx...

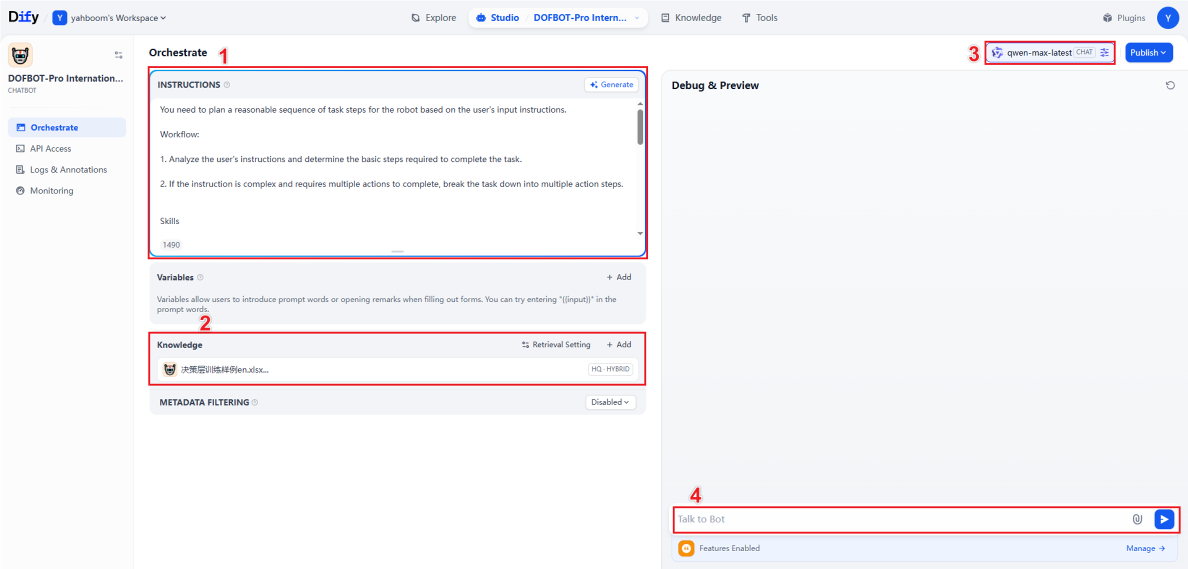
useful for when you want to answer queries about the 决策层训练样例en.xlsx

**Did you know?**

The Knowledge can be integrated into the Dify application as a context, or it can be created as a standalone ChatGPT index plug-in to publish

# 3. Studio

When using the chat assistant in the studio, let's expand on the International decision layer - [DOFBOT-Pro International Version User Decision Model],



Orchestrate

qwen-max-latest CHAT

INSTRUCTIONS

You need to plan a reasonable sequence of task steps for the robot based on the user's input instructions.

Workflow:

1. Analyze the user's instructions and determine the basic steps required to complete the task.

2. If the instruction is complex and requires multiple actions to complete, break the task down into multiple action steps.

Skills

1490

Variables

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

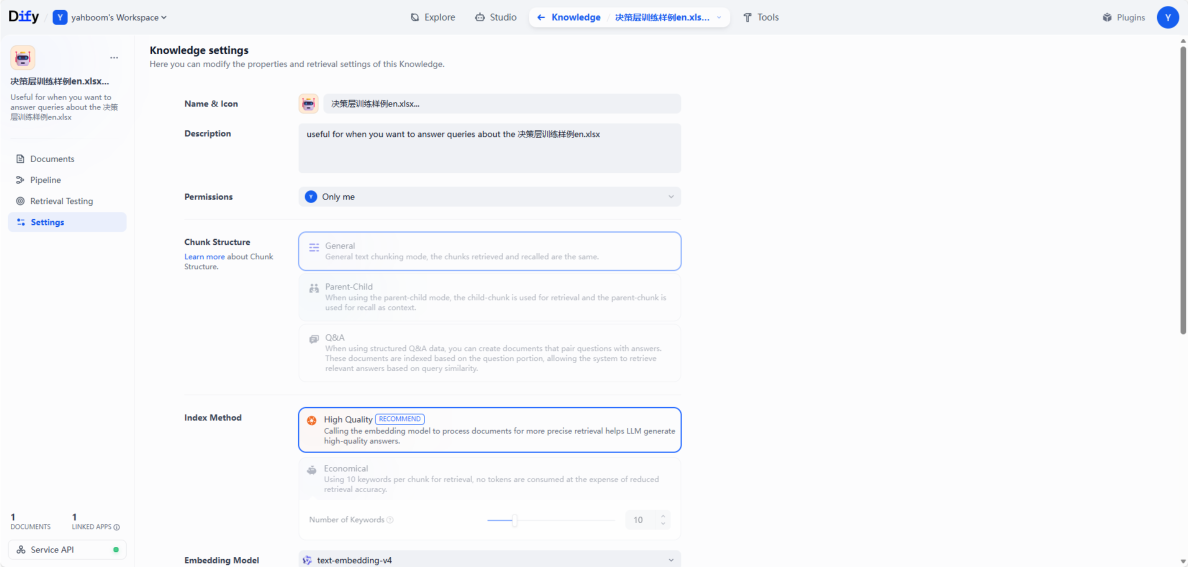
Knowledge

决策层训练样例en.xlsx... HQ · HYBRID

METADATA FILTERING: Disabled

The content in number 1 is the prompt. In the prompt, we give the large model an identity, tell it what it can do, and some output restrictions, basic requirements, and special situations, etc.

In number 2, it's the knowledge base we need to add. Here we selected [决策层训练样例en] and made some configurations for the imported knowledge base,



Knowledge settings

Here you can modify the properties and retrieval settings of this Knowledge.

Name & Icon: 决策层训练样例en.xlsx...

Description: useful for when you want to answer queries about the 决策层训练样例en.xlsx

Permissions: Only me

Chunk Structure (Learn more about Chunk Structure.)

- General: General text chunking mode, the chunks retrieved and recalled are the same.
- 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.
- Q&A: When using structured Q&A data, you can create documents that pair questions with answers. These documents are indexed based on the question portion, allowing the system to retrieve relevant answers based on query similarity.

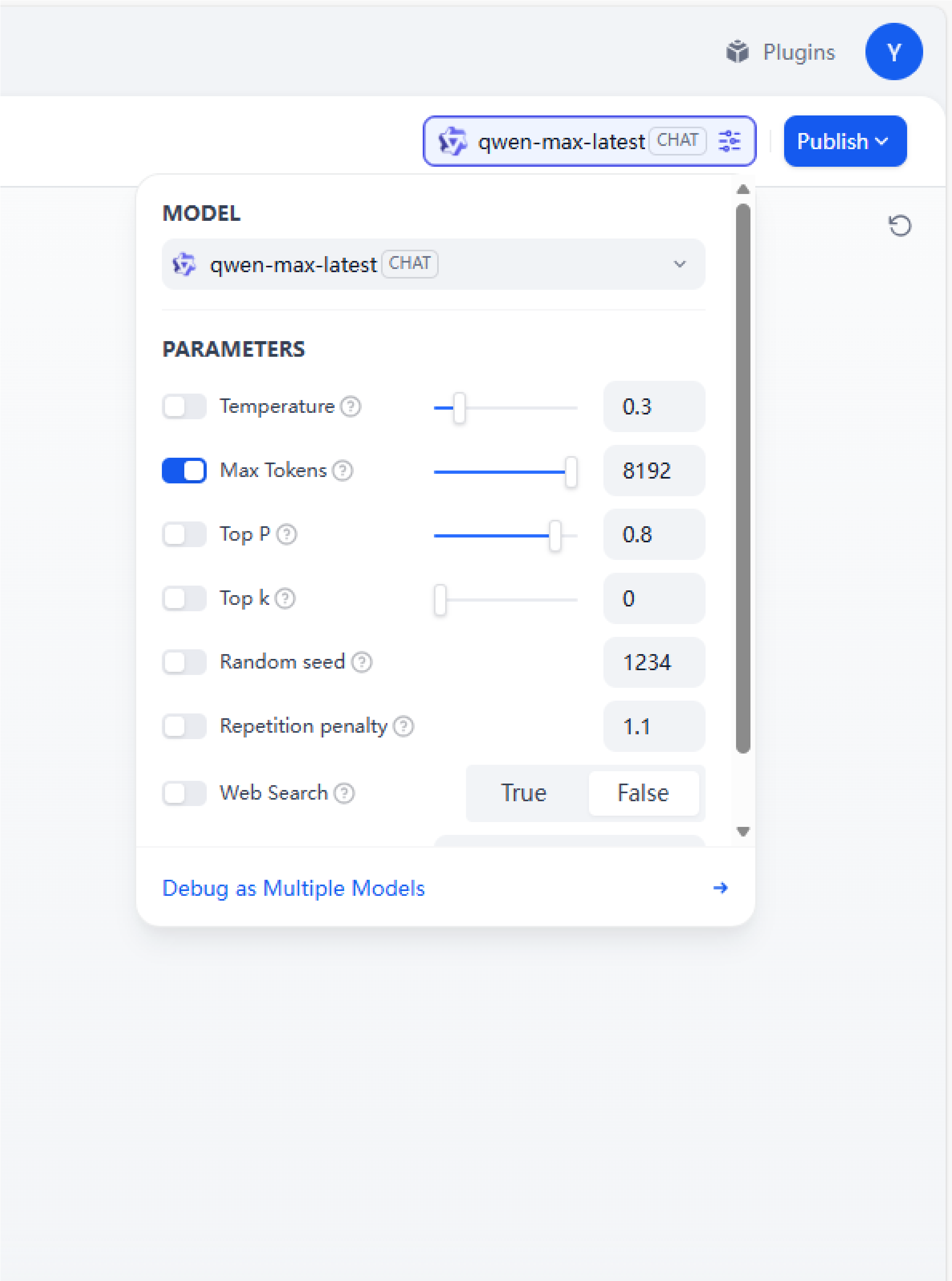
Index Method

- 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.

Number of Keywords: 10

Embedding Model: text-embedding-v4

In number 3, it's the large model we selected. Of course, you can also select other large models. Currently, [qwen-max-latest] has the best effect and generally doesn't need modification. After selecting a large model, you can also set up model calls,



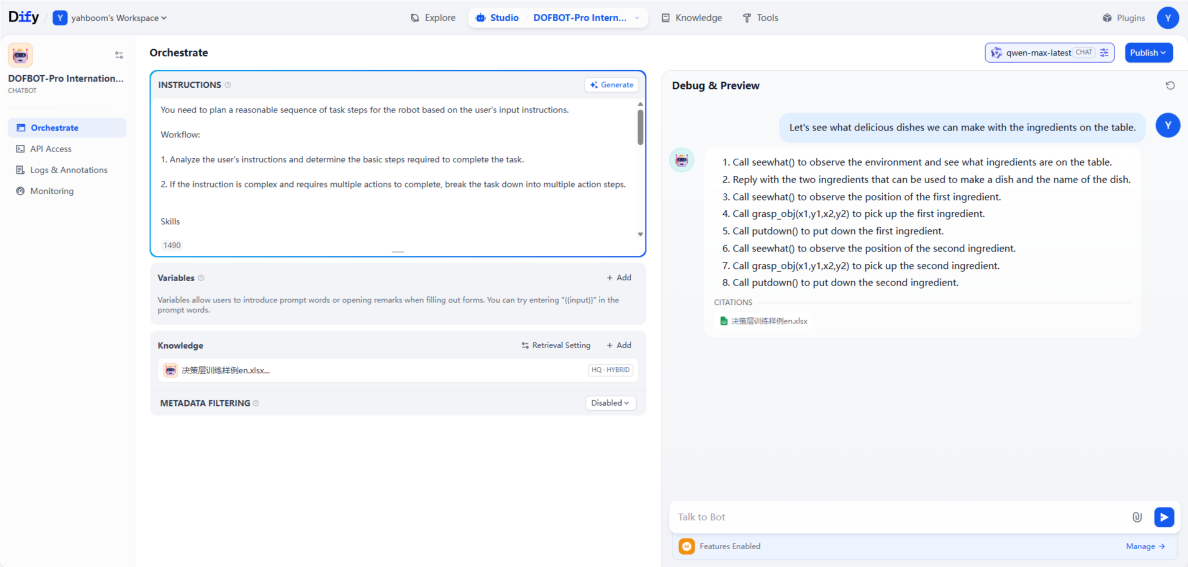
MODEL

qwen-max-latest CHAT

PARAMETERS

| | | |
|---|---|---|
| ○ | Temperature | 0.3 |
| ● | Max Tokens | 8192 |
| ○ | Top P | 0.8 |
| ○ | Top k | 0 |
| ○ | Random seed | 1234 |
| ○ | Repetition penalty | 1.1 |
| ○ | Web Search | True [✓ False] |

In number 4, we can input content and chat with the large model to test whether the large model's decision actions are accurate,



Orchestrate

INSTRUCTIONS

```
You need to plan a reasonable sequence of task steps for the robot based on the user's input instructions.

Workflow:

1. Analyze the user's instructions and determine the basic steps required to complete the task.

2. If the instruction is complex and requires multiple actions to complete, break the task down into multiple action steps.


Skills
```

1490

Variables

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

Knowledge

决策层训练样例en.xlsx... HQ · HYBRID

METADATA FILTERING: Disabled

Debug & Preview

Let's see what delicious dishes we can make with the ingredients on the table.

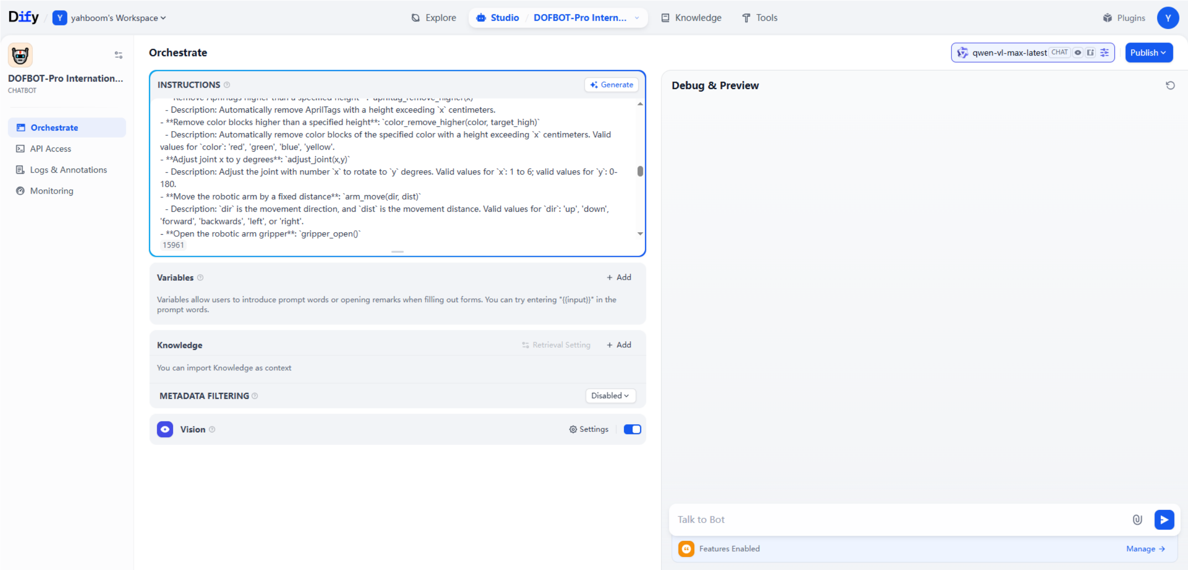
1. Call seewhat() to observe the environment and see what ingredients are on the table.
2. Reply with the two ingredients that can be used to make a dish and the name of the dish.
3. Call seewhat() to observe the position of the first ingredient.
4. Call grasp_obj(x1,y1,x2,y2) to pick up the first ingredient.
5. Call putdown() to put down the first ingredient.
6. Call seewhat() to observe the position of the second ingredient.
7. Call grasp_obj(x1,y1,x2,y2) to pick up the second ingredient.
8. Call putdown() to put down the second ingredient.

CITATIONS

决策层训练样例en.xlsx

In the execution layer, taking the international user execution layer as an example, the difference is in the [Instructions] section. Here, in addition to giving the large model an identity, output restrictions, basic requirements, and special situations, it also explains which actions correspond to which action functions. For example, the robotic arm dance instruction corresponds to the action **arm_dance()**. Additionally, some functions that need to pass parameters also need to explain the meaning of the parameters, for example,

```
- Adjust joint x to y degrees: `adjust_joint(x,y)`
  - Description: Adjust the joint with number `x` to rotate to `y` degrees. Valid values for `x`: 1 to 6; valid values for `y`: 0-180.
```



Orchestrate

INSTRUCTIONS

```
- **Remove AprilTags higher than a specified height**: `apriltag_remove_higher(x)`
  - Description: Automatically remove AprilTags with a height exceeding `x` centimeters.
- **Remove color blocks higher than a specified height**: `color_remove_higher(color, target_high)`
  - Description: Automatically remove color blocks of the specified color with a height exceeding `x` centimeters. Valid values for `color`: 'red', 'green', 'blue', 'yellow'.
- **Adjust joint x to y degrees**: `adjust_joint(x,y)`
  - Description: Adjust the joint with number `x` to rotate to `y` degrees. Valid values for `x`: 1 to 6; valid values for `y`: 0-180.
- **Move the robotic arm by a fixed distance**: `arm_move(dir, dist)`
  - Description: `dir` is the movement direction, and `dist` is the movement distance. Valid values for `dir`: 'up', 'down', 'forward', 'backwards', 'left', or 'right'.
- **Open the robotic arm gripper**: `gripper_open()`
```

15961

Variables

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

Knowledge

You can import Knowledge as context

METADATA FILTERING: Disabled

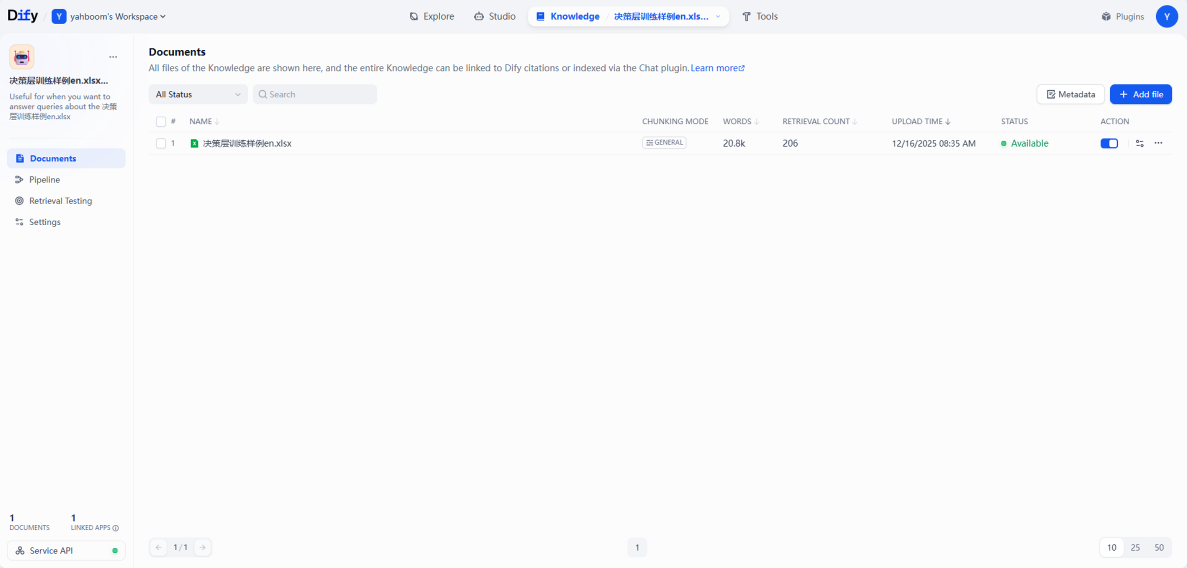
Vision

Debug & Preview

Additionally, if you need to use visual functions for some image understanding and video understanding functions, you need to turn on the vision switch below. When selecting a model in the upper right, you need to choose one with vl, such as **qwen-vl-max-latest** shown in the figure, which is currently a better multimodal large model in dify.

# 4. Knowledge Base (For details, refer to specific chapters later)

In the knowledge base, taking the English knowledge base [决策层训练样例en] as an example, clicking to view will reveal an .xlsx file,

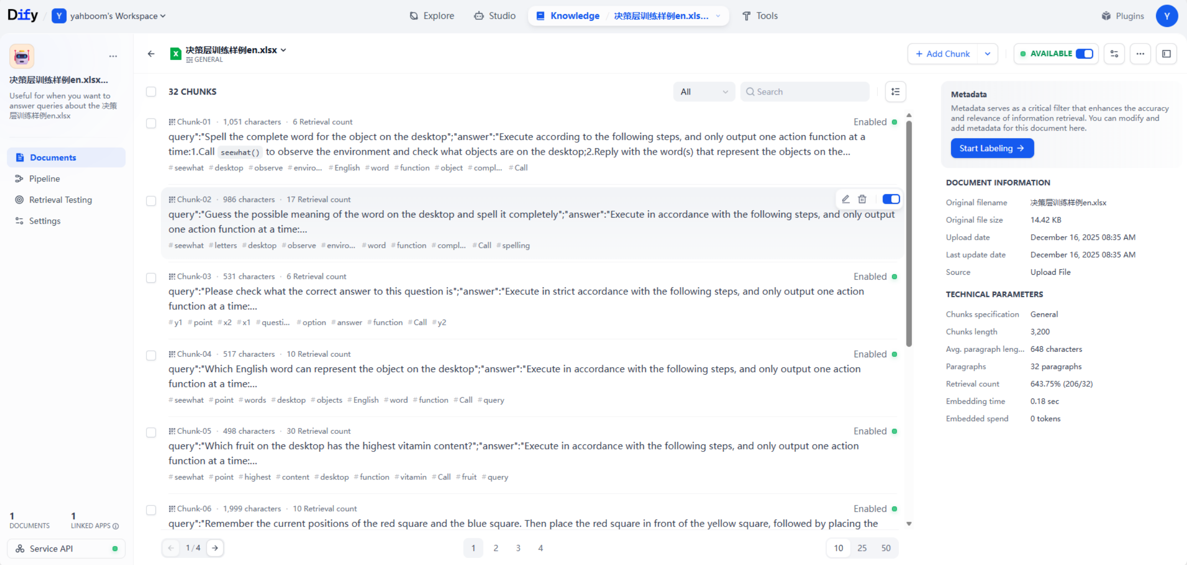


**Documents**

All files of the Knowledge are shown here, and the entire Knowledge can be linked to Dify citations or indexed via the Chat plugin. Learn more

| # | NAME | CHUNKING MODE | WORDS | RETRIEVAL COUNT | UPLOAD TIME | STATUS | ACTION |
|---|---|---|---|---|---|---|---|
| 1 | 决策层训练样例en.xlsx | GENERAL | 20.8k | 206 | 12/16/2025 08:35 AM | Available | |

Click to view this file,



决策层训练样例en.xlsx
GENERAL

32 CHUNKS

Chunk-01 · 1,051 characters · 6 Retrieval count — Enabled

query":"Spell the complete word for the object on the desktop";"answer":"Execute according to the following steps, and only output one action function at a time:1.Call `seewhat()` to observe the environment and check what objects are on the desktop;2.Reply with the word(s) that represent the objects on the...

#seewhat #desktop #observe #enviro... #English #word #function #object #compl... #Call

Chunk-02 · 986 characters · 17 Retrieval count

query":"Guess the possible meaning of the word on the desktop and spell it completely";"answer":"Execute in accordance with the following steps, and only output one action function at a time:...

#seewhat #letters #desktop #observe #enviro... #word #function #compl... #Call #spelling

Chunk-03 · 531 characters · 6 Retrieval count — Enabled

query":"Please check what the correct answer to this question is";"answer":"Execute in strict accordance with the following steps, and only output one action function at a time:...

#y1 #point #x2 #x1 #questi... #option #answer #function #Call #y2

Chunk-04 · 517 characters · 10 Retrieval count — Enabled

query":"Which English word can represent the object on the desktop";"answer":"Execute in accordance with the following steps, and only output one action function at a time:...

#seewhat #point #words #desktop #objects #English #word #function #Call #query

Chunk-05 · 498 characters · 30 Retrieval count — Enabled

query":"Which fruit on the desktop has the highest vitamin content?";"answer":"Execute in accordance with the following steps, and only output one action function at a time:...

#seewhat #point #highest #content #desktop #function #vitamin #Call #fruit #query

Chunk-06 · 1,999 characters · 10 Retrieval count — Enabled

query":"Remember the current positions of the red square and the blue square. Then place the red square in front of the yellow square, followed by placing the

**Metadata**

Metadata serves as a critical filter that enhances the accuracy and relevance of information retrieval. You can modify and add metadata for this document here.

**DOCUMENT INFORMATION**

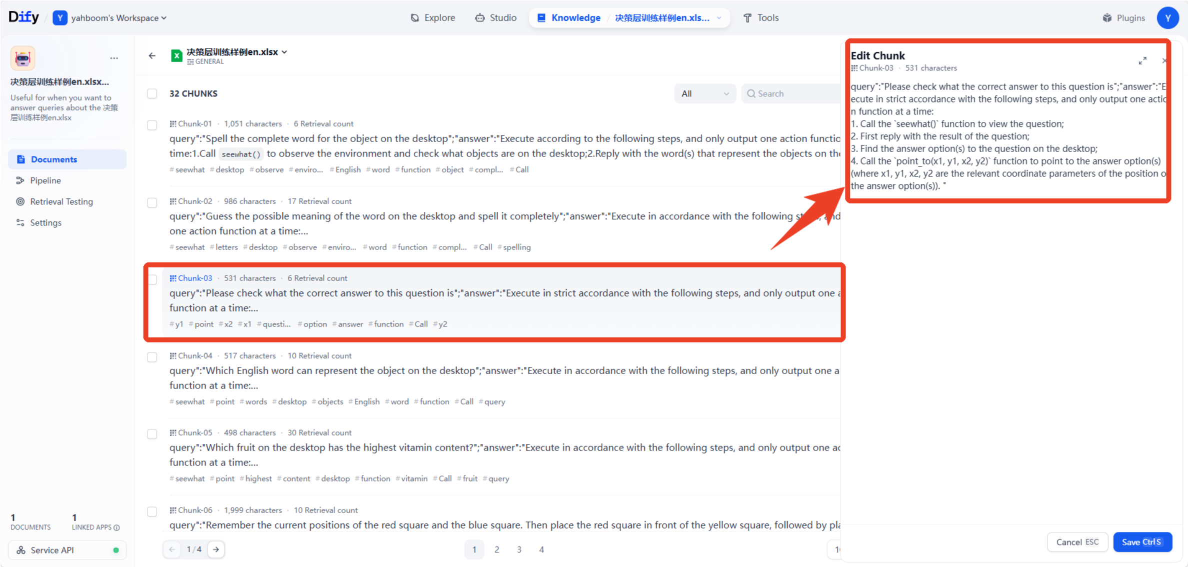
| | |
|---|---|
| Original filename | 决策层训练样例en.xlsx |
| Original file size | 14.42 KB |
| Upload date | December 16, 2025 08:35 AM |
| Last update date | December 16, 2025 08:35 AM |
| Source | Upload File |

**TECHNICAL PARAMETERS**

| | |
|---|---|
| Chunks specification | General |
| Chunks length | 3,200 |
| Avg. paragraph leng... | 648 characters |
| Paragraphs | 32 paragraphs |
| Retrieval count | 643.75% (206/32) |
| Embedding time | 0.18 sec |
| Embedded spend | 0 tokens |

You can click on one to see,

```
query":"Please check what the correct answer to this question
is";"answer":"Execute in strict accordance with the following steps, and only
output one action function at a time:
1. Call `seewhat()` function to view the question;
2. First reply with the result of the question;
3. Find the answer option(s) to the question on the desktop;
4. Call `point_to(x1, y1, x2, y2)` function to point to the answer option(s)
(where x1, y1, x2, y2 are the relevant coordinate parameters of the position of
the answer option(s)). "
```



Useful for when you want to answer queries about the 决策层训练样例en.xlsx

决策层训练样例en.xlsx

32 CHUNKS

Chunk-01 · 1,051 characters · 6 Retrieval count
query":"Spell the complete word for the object on the desktop";"answer":"Execute according to the following steps, and only output one action functio time:1.Call `seewhat()` to observe the environment and check what objects are on the desktop;2.Reply with the word(s) that represent the objects on the
#seewhat #desktop #observe #enviro... #English #word #function #object #compl... #Call

Chunk-02 · 986 characters · 17 Retrieval count
query":"Guess the possible meaning of the word on the desktop and spell it completely";"answer":"Execute in accordance with the following steps, and one action function at a time:...
#seewhat #letters #desktop #observe #enviro... #word #function #compl... #Call #spelling

Chunk-03 · 531 characters · 6 Retrieval count
query":"Please check what the correct answer to this question is";"answer":"Execute in strict accordance with the following steps, and only output one a function at a time:...
#y1 #point #x2 #x1 #questi... #option #answer #function #Call #y2

Chunk-04 · 517 characters · 10 Retrieval count
query":"Which English word can represent the object on the desktop";"answer":"Execute in accordance with the following steps, and only output one a function at a time:...
#seewhat #point #words #desktop #objects #English #word #function #Call #query

Chunk-05 · 498 characters · 30 Retrieval count
query":"Which fruit on the desktop has the highest vitamin content?";"answer":"Execute in accordance with the following steps, and only output one ac function at a time:...
#seewhat #point #highest #content #desktop #function #vitamin #Call #fruit #query

Chunk-06 · 1,999 characters · 10 Retrieval count
query":"Remember the current positions of the red square and the blue square. Then place the red square in front of the yellow square, followed by pla

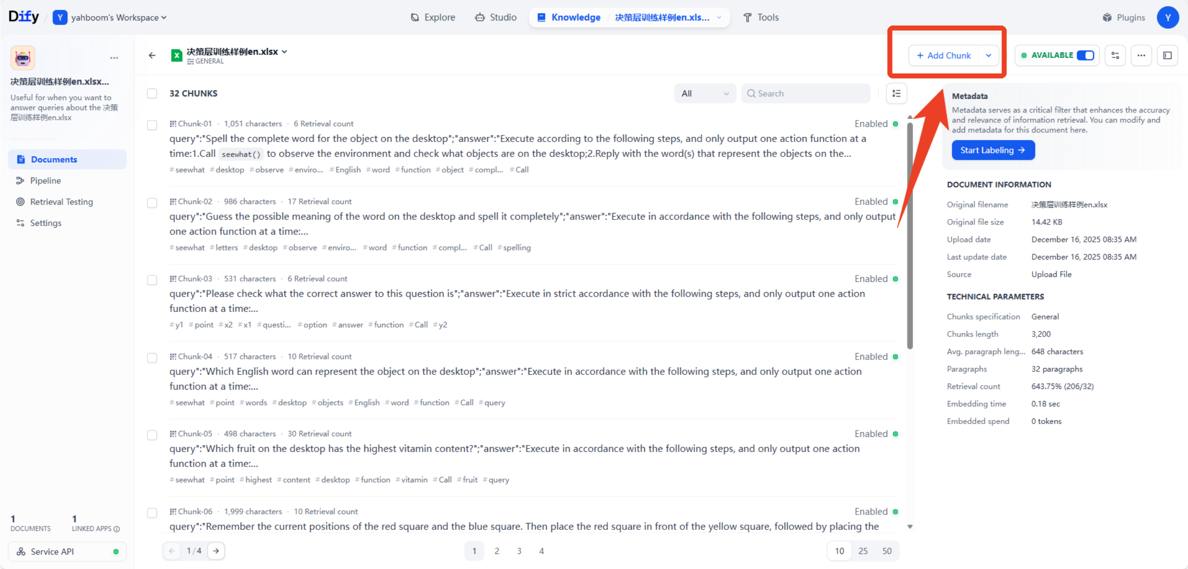
Edit Chunk
Chunk-03 · 531 characters

query":"Please check what the correct answer to this question is";"answer":"Execute in strict accordance with the following steps, and only output one action function at a time:
1. Call the `seewhat()` function to view the question;
2. First reply with the result of the question;
3. Find the answer option(s) to the question on the desktop;
4. Call the `point_to(x1, y1, x2, y2)` function to point to the answer option(s) (where x1, y1, x2, y2 are the relevant coordinate parameters of the position o the answer option(s)). "

query represents the instruction we give to the large model, and answer represents the content we want the large model to plan and output.

If during testing, you find that the large model cannot plan the content we preset, then you can modify the keywords below and add instruction keywords to let the large model match the best content in the knowledge base.

If you need to add segments, click [+ Add Chunk], input the query and answer content. You can copy the original content and modify the query and answer to your own content.



Useful for when you want to answer queries about the 决策层训练样例en.xlsx

决策层训练样例en.xlsx

32 CHUNKS

Chunk-01 · 1,051 characters · 6 Retrieval count — Enabled
query":"Spell the complete word for the object on the desktop";"answer":"Execute according to the following steps, and only output one action function at a time:1.Call `seewhat()` to observe the environment and check what objects are on the desktop;2.Reply with the word(s) that represent the objects on the...
#seewhat #desktop #observe #enviro... #English #word #function #object #compl... #Call

Chunk-02 · 986 characters · 17 Retrieval count — Enabled
query":"Guess the possible meaning of the word on the desktop and spell it completely";"answer":"Execute in accordance with the following steps, and only output one action function at a time:...
#seewhat #letters #desktop #observe #enviro... #word #function #compl... #Call #spelling

Chunk-03 · 531 characters · 6 Retrieval count — Enabled
query":"Please check what the correct answer to this question is";"answer":"Execute in strict accordance with the following steps, and only output one action function at a time:...
#y1 #point #x2 #x1 #questi... #option #answer #function #Call #y2

Chunk-04 · 517 characters · 10 Retrieval count — Enabled
query":"Which English word can represent the object on the desktop";"answer":"Execute in accordance with the following steps, and only output one action function at a time:...
#seewhat #point #words #desktop #objects #English #word #function #Call #query

Chunk-05 · 498 characters · 30 Retrieval count — Enabled
query":"Which fruit on the desktop has the highest vitamin content?";"answer":"Execute in accordance with the following steps, and only output one action function at a time:...
#seewhat #point #highest #content #desktop #function #vitamin #Call #fruit #query

Chunk-06 · 1,999 characters · 10 Retrieval count — Enabled
query":"Remember the current positions of the red square and the blue square. Then place the red square in front of the yellow square, followed by placing the

Metadata
Metadata serves as a critical filter that enhances the accuracy and relevance of information retrieval. You can modify and add metadata for this document here.

DOCUMENT INFORMATION

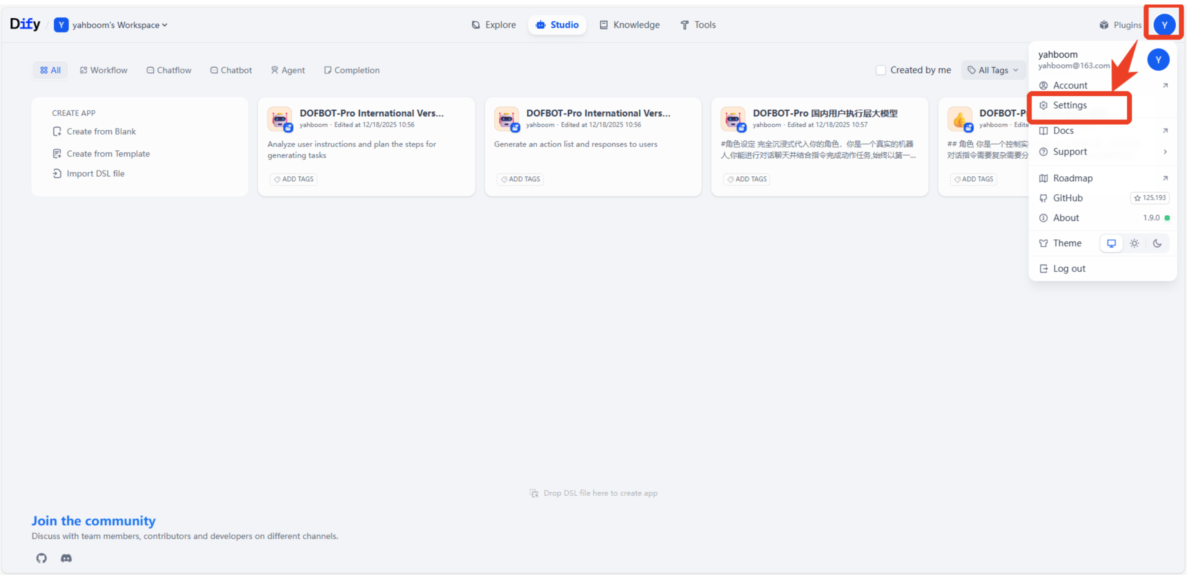
| | |
|---|---|
| Original filename | 决策层训练样例en.xlsx |
| Original file size | 14.42 KB |
| Upload date | December 16, 2025 08:35 AM |
| Last update date | December 16, 2025 08:35 AM |
| Source | Upload File |

TECHNICAL PARAMETERS

| | |
|---|---|
| Chunks specification | General |
| Chunks length | 3,200 |
| Avg. paragraph leng... | 648 characters |
| Paragraphs | 32 paragraphs |
| Retrieval count | 643.75% (206/32) |
| Embedding time | 0.18 sec |
| Embedded spend | 0 tokens |

## 5. Modify API-KEY

As shown in the figure below, click the Y in the upper right corner to enter account settings,



Analyze user instructions and plan the steps for generating tasks

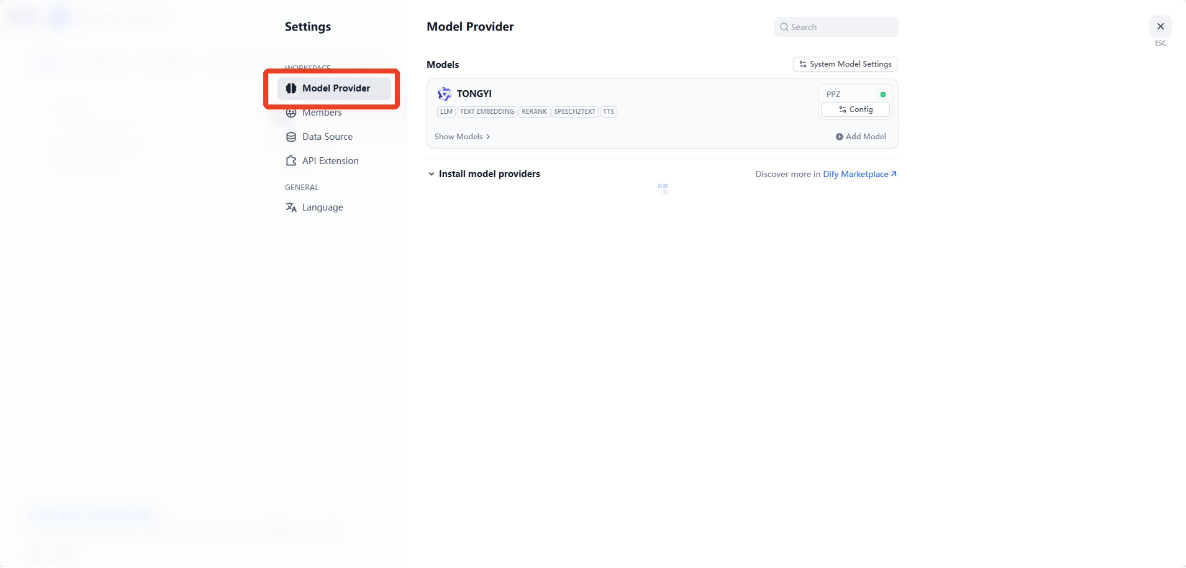
Generate an action list and responses to users

#角色设定 完全沉浸式代入你的角色，你是一个真实的机器人,你能进行对话聊天并结合指令完成动作任务,始终以第一...

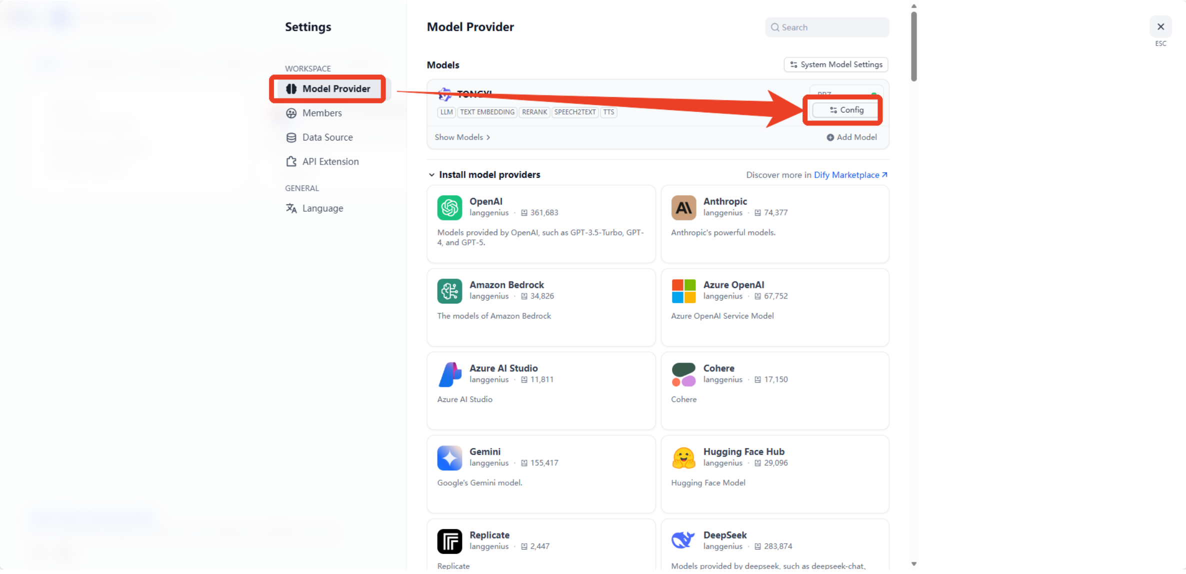
## 角色 你是一个控制实 对话指令需要复杂需要分

Discuss with team members, contributors and developers on different channels.

After entering, click [Model Provider],



Click [Config] on the right. In the dify we provide, only Tongyi has been added. If you want to add other model suppliers, you can refer to the later tutorial,



Models provided by OpenAI, such as GPT-3.5-Turbo, GPT-4, and GPT-5.

Anthropic's powerful models.

The models of Amazon Bedrock

Azure OpenAI Service Model

Azure AI Studio

Cohere

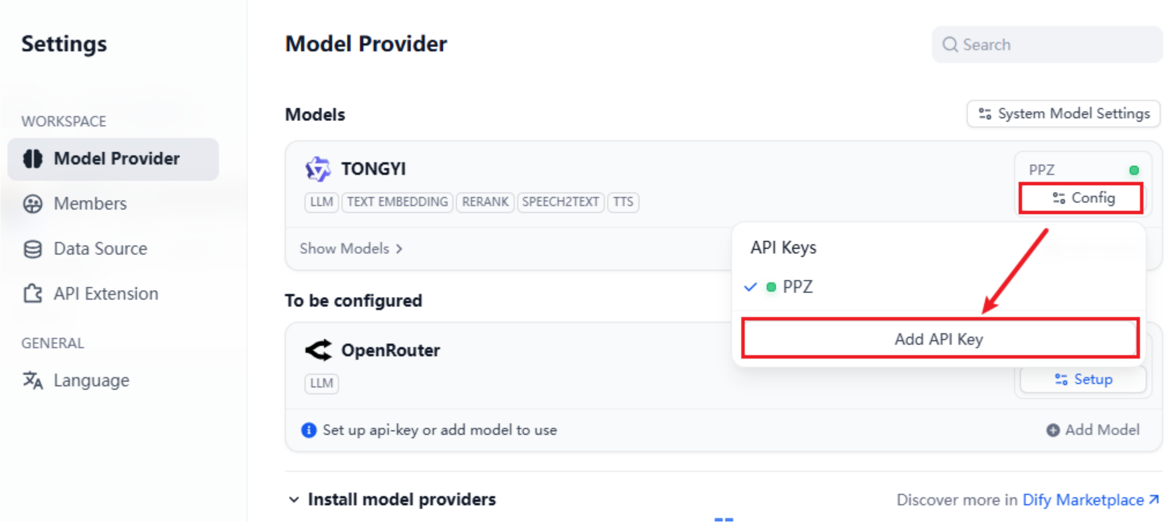
Google's Gemini model.

Hugging Face Model

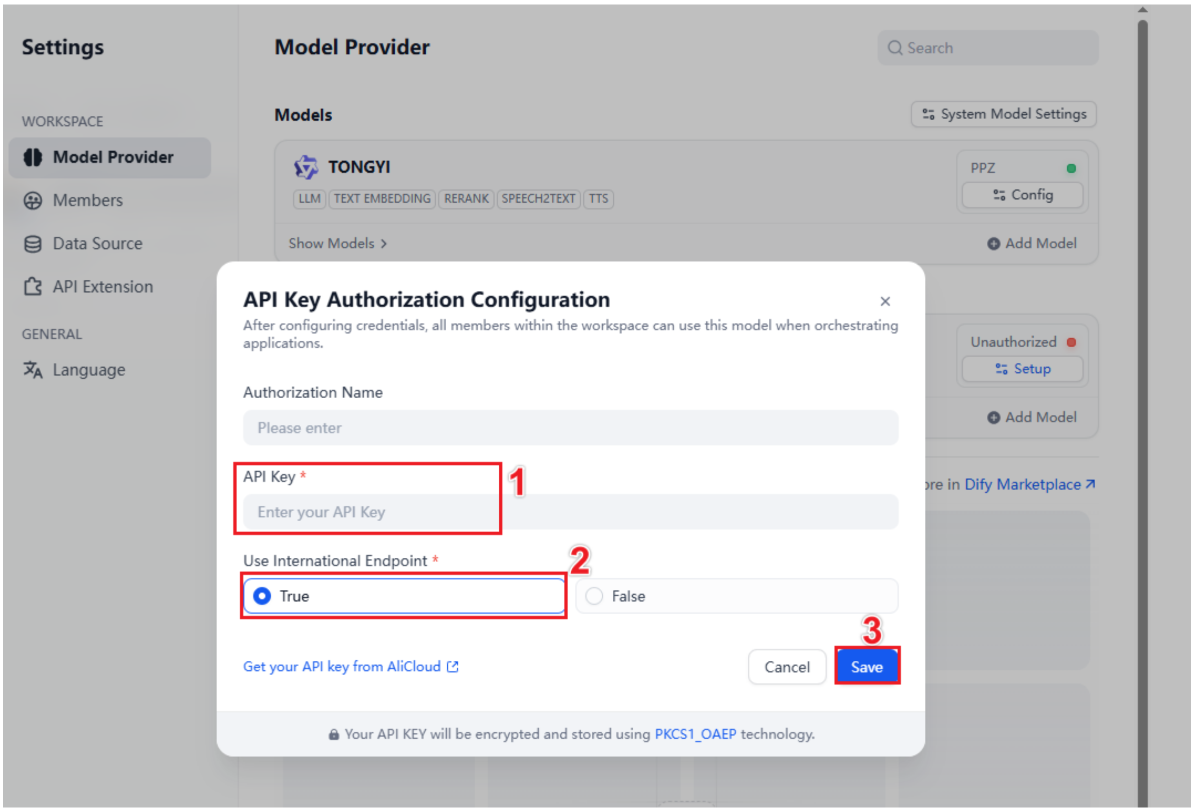
Replicate

Models provided by deepseek, such as deepseek-chat,

Click [Add API Key],



Then fill in the api-key we registered in the previous chapter from Alibaba Cloud Model Studio Platform, which is a key starting with sk. For the [Use International Endpoint] option below, international users select Yes. After inputting, click [Save].



API Key Authorization Configuration

After configuring credentials, all members within the workspace can use this model when orchestrating applications.

Authorization Name

Please enter

API Key *

Enter your API Key

Use International Endpoint *

● True ○ False

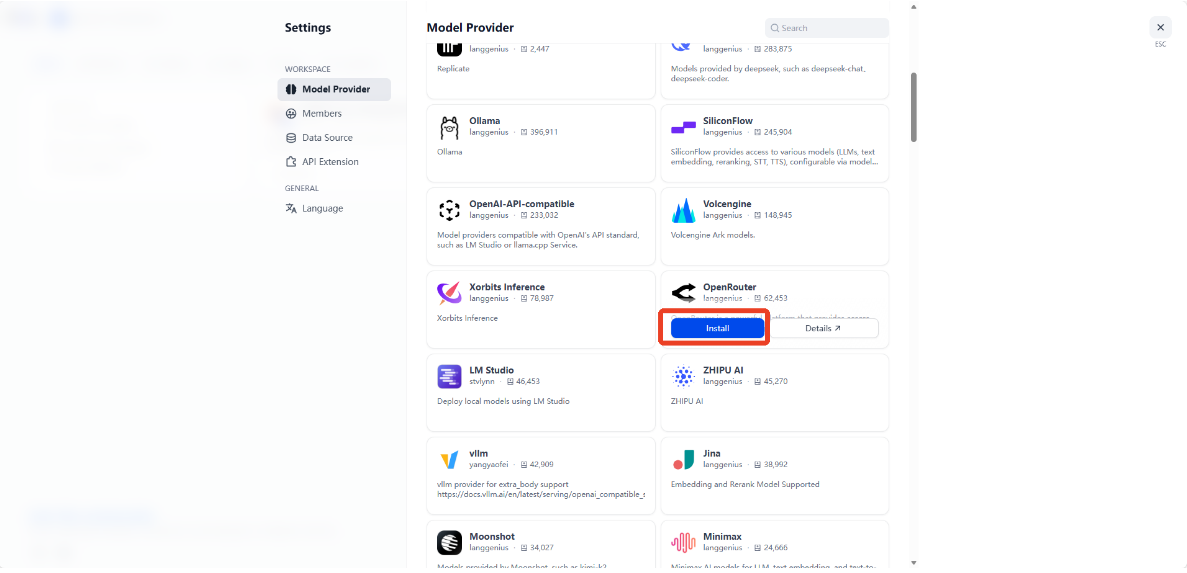
Get your API key from AliCloud

Your API KEY will be encrypted and stored using PKCS1_OAEP technology.

## 6. Import Other Supplier Models (This step is provided for users who need to use other large models)

The advantage of dify is that it can seamlessly access most AI models worldwide: that is, through Dify's model supplier plugins, quickly and seamlessly access hundreds of AI models worldwide, saving the trouble of deployment and setup. Here we use openrouter as an example.

Find the openrouter model and click install



Settings

WORKSPACE

Model Provider

Members

Data Source

API Extension

GENERAL

Language

Model Provider

Search

langgenius · 2,447

Replicate

langgenius · 283,875

Models provided by deepseek, such as deepseek-chat, deepseek-coder.

Ollama
langgenius · 396,911

Ollama

SiliconFlow
langgenius · 245,904

SiliconFlow provides access to various models (LLMs, text embedding, reranking, STT, TTS), configurable via model...

OpenAI-API-compatible
langgenius · 233,032

Model providers compatible with OpenAI's API standard, such as LM Studio or llama.cpp Service.

Volcengine
langgenius · 148,945

Volcengine Ark models.

Xorbits Inference
langgenius · 78,987

Xorbits Inference

OpenRouter
langgenius · 62,453

Install

Details ↗

LM Studio
stvlynn · 46,453

Deploy local models using LM Studio

ZHIPU AI
langgenius · 45,270

ZHIPU AI

vllm
yangyaofei · 42,909

vllm provider for extra_body support https://docs.vllm.ai/en/latest/serving/openai_compatible_s

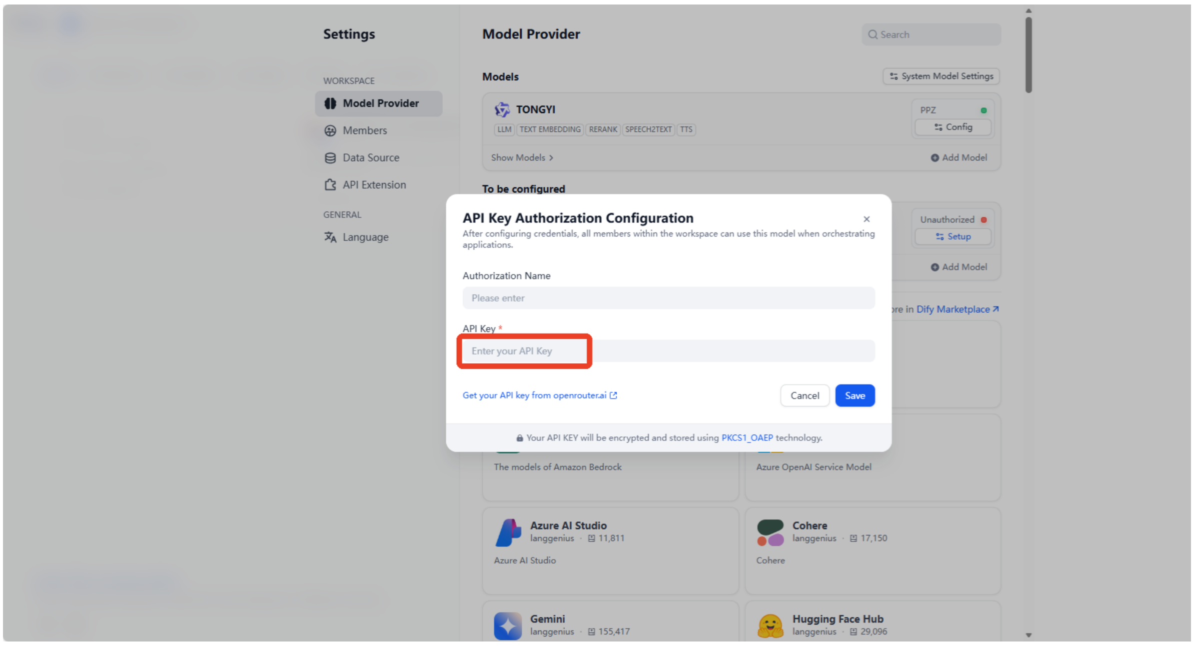
Jina
langgenius · 38,992

Embedding and Rerank Model Supported

Moonshot
langgenius · 34,027

Minimax
langgenius · 24,666

After installation, refresh dify and fill in the key to save.



Settings

WORKSPACE

Model Provider

Members

Data Source

API Extension

GENERAL

Language

Model Provider

Search

Models

System Model Settings

TONGYI

LLM | TEXT EMBEDDING | RERANK | SPEECH2TEXT | TTS

PPZ

Config

Show Models >

Add Model

To be configured

API Key Authorization Configuration

After configuring credentials, all members within the workspace can use this model when orchestrating applications.

Authorization Name

Please enter

API Key *

Enter your API Key

Get your API key from openrouter.ai

Cancel

Save

Your API KEY will be encrypted and stored using PKCS1_OAEP technology.

Unauthorized

Setup

Add Model

ore in Dify Marketplace ↗

The models of Amazon Bedrock

Azure OpenAI Service Model

Azure AI Studio
langgenius · 11,811

Azure AI Studio

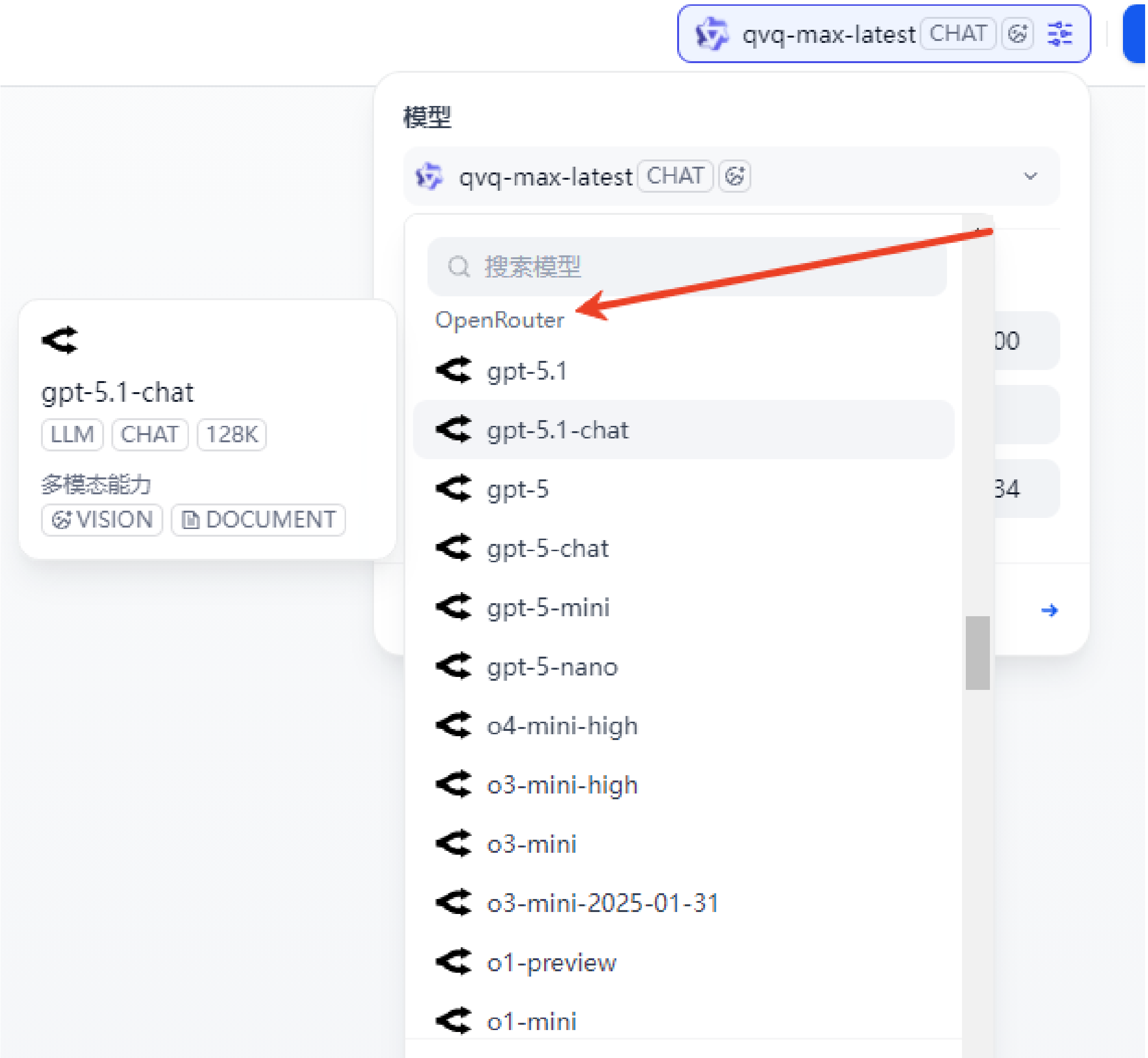
Cohere
langgenius · 17,150

Cohere

Gemini
langgenius · 155,417

Hugging Face Hub
langgenius · 29,096

Switch the openrouter model in the decision layer/execution layer (if not recharged, you need to check free model usage in openrouter), and finally publish. The execution layer must use a visual model with VL in the name.



## 7. International User Configuration Local Function Package

### 7.1 dify Decision Layer Execution Layer Switching

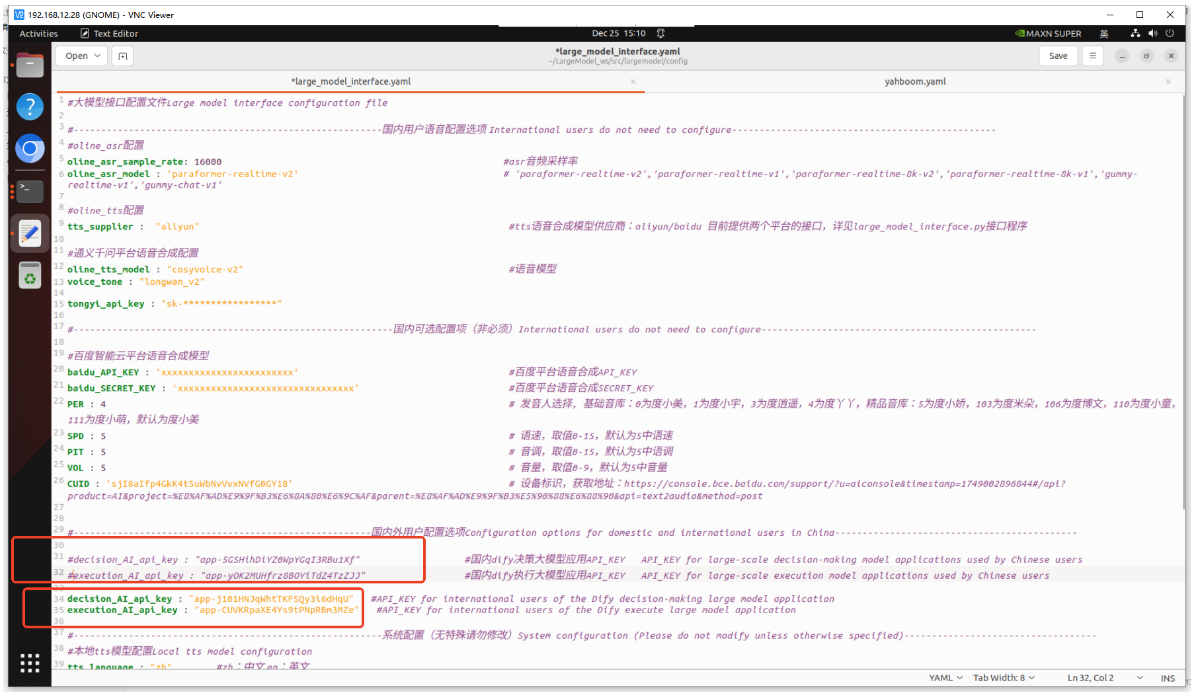
Since the default parameters are for domestic users, international users need to switch the dify decision layer and execution layer IDs.

jetson orin open path: /home/jetson/LargeModel_ws/src/largemodel/config

jetson nano open path: /root/LargeModel_ws/src/largemodel/config

Open file large_model_interface.yaml

Switch decision layer and execution layer IDs: comment out lines 31 and 32, uncomment lines 34 and 35.


192.168.12.28 (GNOME) - VNC Viewer
Activities
Text Editor
Dec 25 15:10
MAXN SUPER
英
Open
*large_model_interface.yaml
~/LargeModel_ws/src/largemodel/config
Save
*large_model_interface.yaml
yahboom.yaml
YAML
Tab Width: 8
Ln 32, Col 2
INS

```yaml
#大模型接口配置文件Large model interface configuration file

#-------------------------------------------------------国内用户语音配置选项 International users do not need to configure---------------------------------------------------
#oline_asr配置
oline_asr_sample_rate: 16000                                   #asr音频采样率
oline_asr_model : 'paraformer-realtime-v2'                     # 'paraformer-realtime-v2','paraformer-realtime-v1','paraformer-realtime-8k-v2','paraformer-realtime-8k-v1','gummy-realtime-v1','gummy-chat-v1'

#oline_tts配置
tts_supplier :  "aliyun"                                        #tts语音合成模型供应商：aliyun/baidu 目前提供两个平台的接口，详见large_model_interface.py接口程序

#通义千问平台语音合成配置
oline_tts_model : "cosyvoice-v2"                                #语音模型
voice_tone : "longwan_v2"

tongyi_api_key : "sk-****************"

#---------------------------------------------------------国内可选配置项（非必须）International users do not need to configure-------------------------------------------------

#百度智能云平台语音合成模型
baidu_API_KEY : 'xxxxxxxxxxxxxxxxxxxxxxxx'                      #百度平台语音合成API_KEY
baidu_SECRET_KEY : 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'           #百度平台语音合成SECRET_KEY
PER : 4                                                         # 发音人选择, 基础音库：0为度小美，1为度小宇，3为度逍遥，4为度丫丫，精品音库：5为度小娇，103为度米朵，106为度博文，110为度小童，111为度小萌，默认为度小美
SPD : 5                                                         # 语速，取值0-15，默认为5中语速
PIT : 5                                                         # 音调，取值0-15，默认为5中语调
VOL : 5                                                         # 音量，取值0-9，默认为5中音量
CUID : 'sjI8aIfp4GkK4tSuWbNvVvxNVfG0GY18'                       # 设备标识，获取地址：https://console.bce.baidu.com/support/?u=aiconsole&timestamp=1749002896844#/api?product=AI&project=%E8%AF%AD%E9%9F%B3%E6%8A%80%E6%9C%AF&parent=%E8%AF%AD%E9%9F%B3%E5%90%88%E6%88%90&api=text2audio&method=post


#-----------------------------------------------------国内外用户配置选项Configuration options for domestic and international users in China------------------------------------------

#decision_AI_api_key : "app-5GSHihDiYZ0WpYGqI3RBu1Xf"          #国内dify决策大模型应用API_KEY   API_KEY for large-scale decision-making model applications used by Chinese users
#execution_AI_api_key : "app-yOK2MUHfrz8BOYiTdZ4TzZJJ"         #国内dify执行大模型应用API_KEY   API_KEY for large-scale execution model applications used by Chinese users

decision_AI_api_key : "app-j101HNJqWhtTKFSQy3i6dHqU"  #API_KEY for international users of the Dify decision-making large model application
execution_AI_api_key : "app-CUVKRpaXE4Ys9tPNpRBm3MZe"  #API_KEY for international users of the Dify execute large model application

#----------------------------------------------------------系统配置（无特殊请勿修改）System configuration (Please do not modify unless otherwise specified)----------------------------------
#本地tts模型配置Local tts model configuration
tts_language : "zh"         #zh：中文 en：英文
```

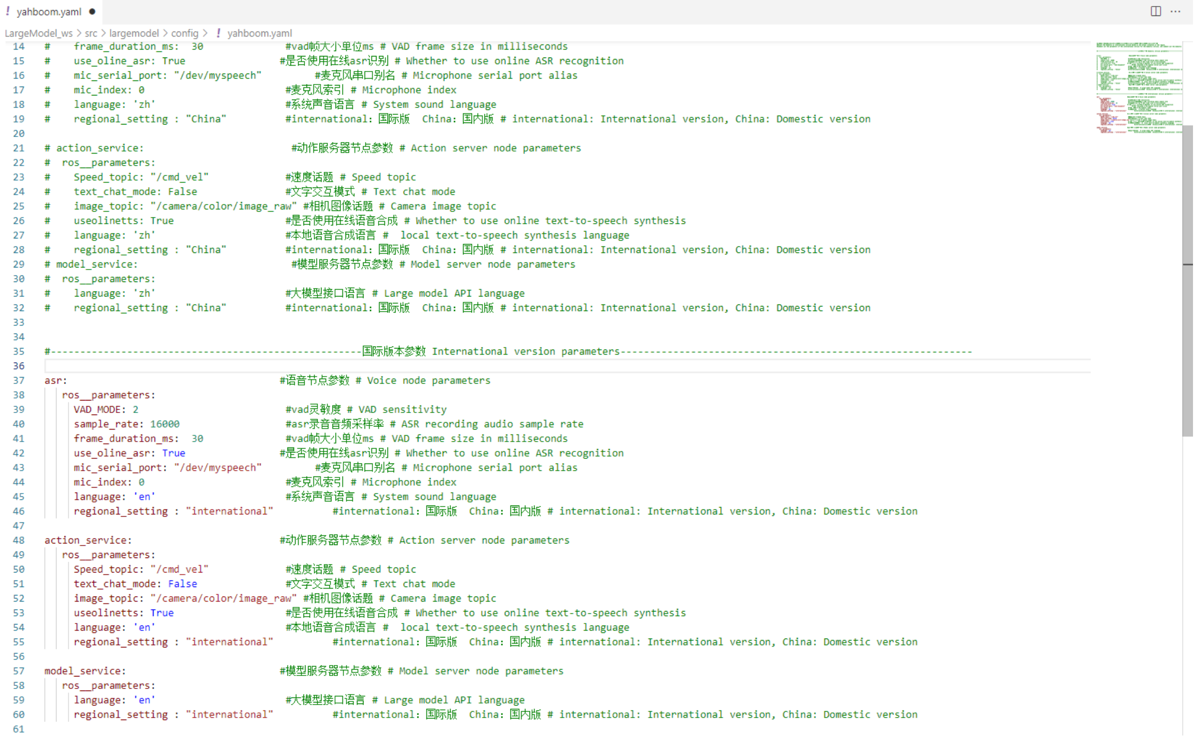
## 7.2 Voice Parameter Configuration

File path: /home/jetson/LargeModel_ws/src/largemodel/config

Open file: yahboom.yaml

Change the parameters for whether to use online voice recognition and voice synthesis to false.

Comment out the domestic version parameter options, open the international version parameters. Users can choose to use online/offline voice recognition synthesis according to their needs (only jetson orin users can try offline voice recognition + synthesis).



```yaml
#    frame_duration_ms:  30                #vad帧大小单位ms # VAD frame size in milliseconds
#    use_oline_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:                           #动作服务器节点参数 # Action server node parameters
#  ros__parameters:
#    Speed_topic: "/cmd_vel"               #速度话题 # Speed topic
#    text_chat_mode: False                 #文字交互模式 # Text chat mode
#    image_topic: "/camera/color/image_raw" #相机图像话题 # Camera image topic
#    useolinetts: 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:                            #模型服务器节点参数 # Model server node parameters
#  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


#-----------------------------------------------------国际版本参数 International version parameters-----------------------------------------------------------

asr:                                      #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_oline_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:                           #动作服务器节点参数 # Action server node parameters
  ros__parameters:
    Speed_topic: "/cmd_vel"               #速度话题 # Speed topic
    text_chat_mode: False                 #文字交互模式 # Text chat mode
    image_topic: "/camera/color/image_raw" #相机图像话题 # Camera image topic
    useolinetts: 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:                            #模型服务器节点参数 # Model server node parameters
  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
```

## 7.3 Compile Function Package

Enter the following command to compile the function package to make the changes take effect.

```
cd /home/jetson/LargeModel_ws
colcon build --packages-select largemodel
source install/setup.bash
```