

# Deploying a Local RAG Knowledge Base

---

## Deploying a Local RAG Knowledge Base

1. Course Content
2. Starting the Dify Service
3. Viewing the Preset Knowledge Base
4. Expanding the RAG Knowledge Base
  - 4.1 Economic Mode Knowledge Base
  - 4.2 High-Quality Mode Knowledge Base
  - 4.3 Recall Test

## 1. Course Content

---

- Master the process and methods for local deployment, debugging, and testing of the RAG knowledge base.
- Master the method for extending the RAG knowledge base based on your specific task scenarios.

[!TIP]

- The RAG knowledge base helps general AI large models provide reference knowledge in vertical domains, preventing AI large models from generating hallucinatory responses and increasing the model's ability to respond with knowledge in vertical domains.
- The RAG knowledge base can help robots quickly expand their generalization capabilities in different task scenarios.

## 2. Starting the Dify Service

---

- Connect to the vehicle's computer via VNC or SSH, and enter the following command in the terminal:

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

```
jetson@yahboom: ~
ROS_DOMAIN_ID: 62 | ROS: humble
my_robot_type: M1 | my_lidar: cl | my_camera: usb
-----
jetson@yahboom: $ sh ~/bringup_dify.sh
WARN[0000] The "DB_USERNAME" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_DATABASE" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_PASSWORD" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_PASSWORD" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_DATABASE" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_USERNAME" variable is not set. Defaulting to a blank string.
WARN[0000] The "CERTBOT_EMAIL" variable is not set. Defaulting to a blank string
.
WARN[0000] The "CERTBOT_DOMAIN" variable is not set. Defaulting to a blank strin
g.
WARN[0000] The "DB_DATABASE" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_PASSWORD" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_USERNAME" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_USERNAME" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_DATABASE" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_PASSWORD" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_DATABASE" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_PASSWORD" variable is not set. Defaulting to a blank string.
WARN[0000] The "DB_USERNAME" variable is not set. Defaulting to a blank string.
[+] Running 10/10
✓ Container docker-sandbox-1      Runni...          0.0s
✓ Container docker-web-1           Running          0.0s
✓ Container docker-ssrf_proxy-1   Ru...          0.0s
✓ Container docker-db-1           Healthy         0.5s
✓ Container docker-redis-1        Running          0.0s
✓ Container docker-worker_beat-1  R...          0.0s
✓ Container docker-plugin_daemon-1 Running          0.0s
✓ Container docker-api-1          Running          0.0s
✓ Container docker-worker-1       Runnin...          0.0s
✓ Container docker-nginx-1        Running          0.0s
jetson@yahboom: $
```

- Check the vehicle's IP address (you can view it on the OLED screen, using `ifconfig`, or directly in the terminal). Enter the vehicle's IP address directly in the browser's address bar to access the Dify management page.

```
jetson@yahboom: ~
jetson@yahboom: ~ 80x24
[System Information]
-----
IP_Address_1: 192.168.2.49
IP_Address_2: 192.168.12.34
MACHINE: OrinNx | ROS_DISTRO: humble | ROS_DOMAIN_ID: 30
ROBOT_TYPE: ROSMASTER-M3Pro | CAMERA_TYPE: dabai_dcw2 | RADAR: Tmini-plus*2
-----
jetson@yahboom: ~$
```

### 3. Viewing the Preset Knowledge Base

- Click on the Knowledge Base page on the homepage. Dify comes pre-configured with two RAG knowledge bases, with the same content but different languages.

[!TIP]

- The preset knowledge base provides training examples for some task scenarios, helping the AI model quickly master relevant skills.

The screenshot shows the Dify Knowledge interface. At the top, there are tabs for Explore, Studio, Knowledge (which is selected), and Tools. Below the tabs, there's a search bar with filters for 'All Knowledge' and 'External Knowledge API'. A red box highlights a specific knowledge base entry titled '决策层训练样例en.xlsx...'. This entry has a thumbnail, a file size of 1.21 MB, and a note stating it's useful for answering queries about the decision layer training examples. It was updated 22 days ago. Below this, there's a 'Did you know?' section with a note about integrating knowledge into the application.

- Clicking on a [决策层训练样例库en] knowledge base reveals a preset file called **Decision Layer Example Library**, which contains the following:
- Decision Layer Sample Library: Stores some preset reference examples related to specific task scenarios.

The screenshot shows the 'Documents' section of the Dify Knowledge interface. On the left, there's a sidebar with options for Documents, Pipeline, Retrieval Testing, and Settings. The main area displays a table of documents. The first document listed is '决策层训练样例en.xlsx', which is described as a preset file for decision layer training examples. The table includes columns for NAME, CHUNKING MODE, WORDS, RETRIEVAL COUNT, UPLOAD TIME, STATUS, and ACTION. The document has a status of 'Available' and was uploaded on 12/16/2025 at 08:35 AM. At the bottom of the page, there are statistics for DOCUMENTS (1) and LINKED APPS (1), and a navigation bar with buttons for Service API, back, forward, and page numbers (1/1).

## 4. Expanding the RAG Knowledge Base

- If you need to expand the knowledge base, click "Create Knowledge Base".

The screenshot shows a modal window for creating a knowledge base. It has three main options: '+ Create Knowledge' (highlighted with a red box), 'Create from Knowledge Pipeline', and 'Connect to an External Knowledge Base'.

- Here, we'll use importing local data as an example.
- Click "Import from file" -> Browse -> Next

- Then, you'll enter the knowledge base configuration page. Click the preview block to view the file chunking effect. Here, select "Economic" for the indexing mode.

[!TIP]

- For beginners, it is recommended to use the economic mode for learning and testing. The difference between the two indexing modes:
  - Economic: Retrieves content from the knowledge base using **keywords**. It cannot perform extended retrieval of similar semantics, and the method of retrieving knowledge fragments is relatively rigid.
  - High-Quality Mode: Requires an embedding model to consume extra tokens and a rerank model, which can achieve more accurate retrieval of similar semantic fragments.
- The default knowledge base mode is the high-quality mode.

## 4.1 Economic Mode Knowledge Base

- After selecting the following configuration, click "Save & Process".

- Then wait for the embedding to complete, and click to ""Go to the document."

The screenshot shows the Dify Knowledge interface. At the top, there are tabs for Explore, Studio, Knowledge (selected), and Tools. Below the tabs, it says 'STEP 3 EXECUTE & FINISH'. On the left, there's a sidebar with 'KNOWLEDGE' and a 'Knowledge created' section. In the center, there's a 'EMBEDDING COMPLETED' section with a file named 'Sample training for the decision-making level.xlsx'. Below this, there are settings for Chunking, Text Preprocessing Rules, Index Method, and Retrieval Setting. A red box highlights the 'Go to document' button at the bottom of this section.

- When the knowledge base is functioning normally, the status will show as available. Then click on the knowledge base file.

The screenshot shows the Dify Documents interface. On the left, there's a sidebar with 'Documents', 'Pipeline', 'Retrieval Testing', and 'Settings'. The main area shows a table of documents. One row is highlighted with a red box, showing the file 'Sample training for the decision-making level.xlsx'. To the right of the file name, the status is shown as 'Available' with a green dot, also highlighted by a red box.

- Afterwards, you can see the segmented knowledge base fragments. The small text below each segment shows the automatically generated keywords for that segment (only available in economic mode).

The screenshot shows the Dify Knowledge interface with the file 'Sample training for the decision-making level.xlsx' selected. The left sidebar shows 'Documents', 'Pipeline', 'Retrieval Testing', and 'Settings'. The main area displays 16 chunks of the document. Each chunk has a checkbox and a preview of its content. A red box highlights the edit icon (pencil) next to the first chunk's preview. To the right, there are sections for 'Metadata' (with a note about enhancing retrieval accuracy), 'DOCUMENT INFORMATION' (listing original filename, file size, upload date, etc.), and 'TECHNICAL PARAMETERS' (listing chunks specification, length, avg. paragraph length, etc.).

- If the keywords do not accurately describe the knowledge fragment, click "Edit" on the right side of the fragment to edit the content or keywords of that fragment. The image below

shows the modified keywords, then click save.

The screenshot shows the Dify interface with a document titled "Sample training for the decision-making level.xlsx". In the center, there's a list of chunks with their content and character count. On the right, an "Edit Chunk" dialog is open for the first chunk, which contains the query "Remove/take away the machine code with a height higher than x cm.", answer "1. Call the function to remove machine code of a specified height x...", and some code. Below the dialog, a "KEYWORDS" section lists "cm x | higher x | query x | away x | Remove x | take x | code x | machine x | height x | answer x | + Add keyword". At the bottom right of the dialog is a red-bordered "Save & Process" button.

## 4.2 High-Quality Mode Knowledge Base

- If you need to use a high-quality knowledge base later, refer to this section of the tutorial.
- The knowledge base creation and file import process is the same as before.
- Here, select "High Quality" for the indexing method, and choose any retrieval method. Here, we use hybrid retrieval as an example. Finally, save and process.

The screenshot shows the "Knowledge" settings page. Under "Index Method", "High Quality" is selected (RECOMMENDED). Under "Retrieval Setting", "Hybrid Search" is selected (RECOMMENDED). The "Save & Process" button at the bottom right is highlighted with a red border.

## 4.3 Recall Test

- Recall testing tests the actual effectiveness of retrieving relevant knowledge snippets from the knowledge base based on the input, helping to optimize the AI model's response performance.
- After opening a knowledge base, click on "Recall Test" on the left side.

Dify yahboom's Workspace Explore Studio Knowledge Sample training fo... Tools Plugins Y

### 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!

All Status   NAME  ID  RETRIEVAL COUNT  UPLOAD TIME  STATUS  ACTION  Metadata

	NAME	CHUNKING MODE	WORDS	RETRIEVAL COUNT	UPLOAD TIME	STATUS	ACTION
<input type="checkbox"/>	1 Sample training for the decision-making level.xlsx	GENERAL	4.3k	0	01/12/2026 04:18 AM	Available	<input type="button" value="Edit"/> <input type="button" value="Delete"/> <input type="button" value="More"/>

1 DOCUMENTS 0 LINKED APPS  1 / 1 10 25 50

- Enter the test content in the source text (simulating user input during actual use), and then click "Test".
- The retrieved paragraphs and the knowledge base related to the input content will appear on the right. The knowledge base tested here is the economic model knowledge base, which retrieves information based on keywords.

Dify yahboom's Workspace Explore Studio Knowledge Sample training fo... Tools Plugins Y

### Retrieval Test

Test the hitting effect of the Knowledge based on the given query text.

SOURCE TEXT  INVERTED INDEX

2 Retrieved Chunks

Chunk-01 173 characters  
query:"Remove/take away the machine code with a height higher than x cm.";"answer":1. Call the function to remove machine code of a specified height with the parameter x."...  
#cm = higher #query #away #Remove #take #code #machin... #height #answer

OPEN ↗

Sample training for the decision-making level.xlsx

Chunk-03 122 characters  
query:"Sort/pick up/grasp machine code No.x.";"answer":1. Call the function to sort machine code with the parameter x."...  
#grasp #query #functions #code #pick #machin... #Call #answer #Sort #sort

OPEN ↗

Sample training for the decision-making level.xlsx

Records

SOURCE	TEXT	TIME
Retrieval Test	remove machine code	01/12/2026 04:23 AM

1 DOCUMENTS 0 LINKED APPS  1 / 200

- If it's a high-quality mode knowledge base, the retrieved snippets will have a SCORE rating. A higher score indicates a higher relevance between the snippet and the input content. High-quality mode knowledge bases can perform associative retrieval of similar semantics, but also consume tokens.

**Dify** **yahoom's Workspace**

**Retrieval Test**  
Test the hitting effect of the Knowledge based on the given query text.

**SOURCE TEXT**

remove machine code

19 / 200

**3 Retrieved Chunks**

!!Chunk-01 - 173 characters   
query:"Remove/take away the machine code with a height higher than x cm.;" answer:"1. Call the function to remove machine code of a specified height with the parameter x. ..."

!!Chunk-03 - 122 characters   
query:"Sort/pick up/grasp machine code No. x;" answer:"1. Call the function to sort machine code with the parameter X. ..."

!!Chunk-04 - 185 characters   
query:"Remove/take away n - colored squares with a height higher than x cm.;" answer:"1. Call the function to remove color blocks of a specified height with the parameters 'n' and x. ..."

**Records**

SOURCE	TEXT	TIME
@ Retrieval Test	remove machine code	01/12/2026 04:25 AM

1 DOCUMENTS 0 LINKED APPS